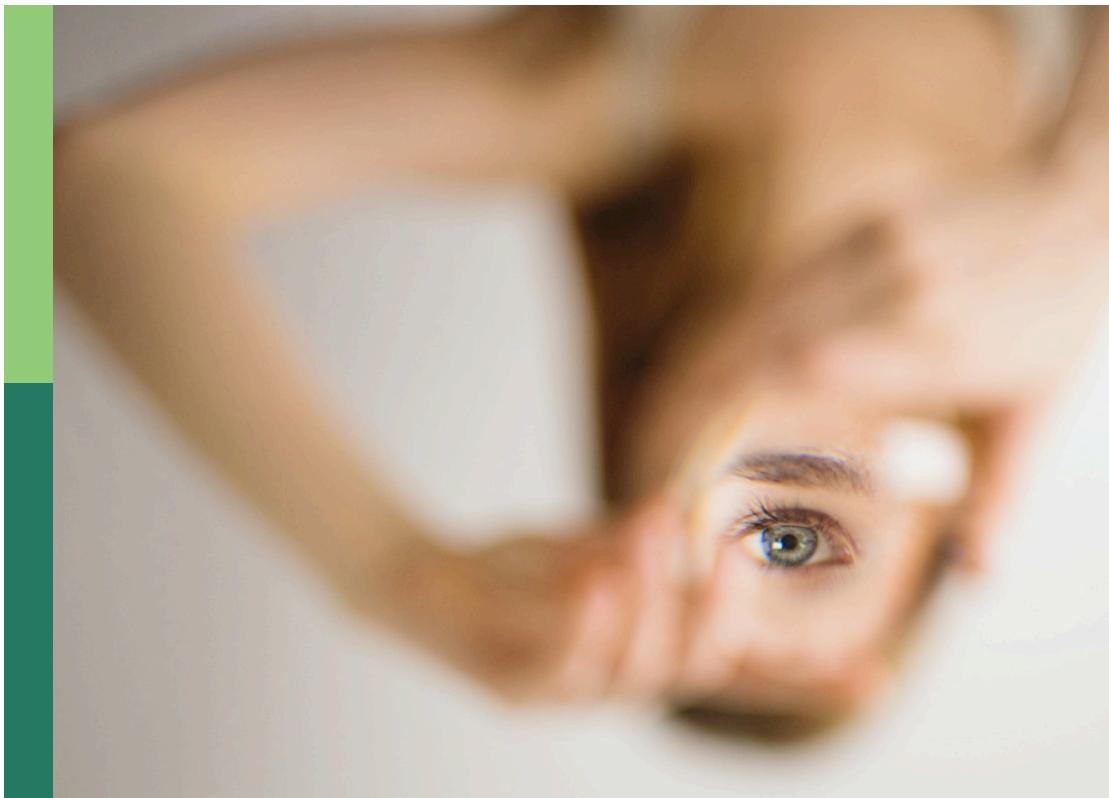
frontiers | Research Topics

# Critical debates on quantitative psychology and measurement: revived and novel perspectives on fundamental problems

**Edited by**
Jana Uher, Jan Ketil Arnulf and Barbara Hanfstingl

**Published in**
Frontiers in Psychology



ISSN 1664-8714
ISBN 978-2-8325-7082-1
DOI 10.3389/978-2-8325-7082-1

**Generative AI statement**
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Critical debates on quantitative psychology and measurement: revived and novel perspectives on fundamental problems

**Topic editors**

Jana Uher — University of Greenwich, United Kingdom

Jan Ketil Arnulf — BI Norwegian Business School, Norway

Barbara Hanfstingl — University of Klagenfurt, Austria

**Citation**

Uher, J., Arnulf, J. K., Hanfstingl, B., eds. (2025). *Critical debates on quantitative psychology and measurement: revived and novel perspectives on fundamental problems*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-7082-1

# Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 16 October 2025
DOI 10.3389/fpsyg.2025.1661765



OPEN ACCESS

EDITED AND REVIEWED BY
Pietro Cipresso,
University of Turin, Italy

*CORRESPONDENCE
Jana Uher
✉ mail@janauher.com

RECEIVED 08 July 2025
ACCEPTED 29 September 2025
PUBLISHED 16 October 2025

CITATION
Uher J, Arnulf JK and Hanfstingl B (2025) Editorial: Critical debates on quantitative psychology and measurement: Revived and novel perspectives on fundamental problems. *Front. Psychol.* 16:1661765.
doi: 10.3389/fpsyg.2025.1661765



# Editorial: Critical debates on quantitative psychology and measurement: Revived and novel perspectives on fundamental problems

Jana Uher [1]*, Jan Ketil Arnulf [2,3] and Barbara Hanfstingl [4]

[1]School of Human Sciences, University of Greenwich, London, United Kingdom, [2]BI Norwegian Business School, Oslo, Norway, [3]Norwegian Defence University College, Oslo, Norway, [4]Department of Psychology, University of Klagenfurt, Klagenfurt, Austria

KEYWORDS

**measurement, quantitative psychology, psychometrics, language models, ontology, epistemology, methodology, semantics**

Editorial on the Research Topic
Critical debates on quantitative psychology and measurement: Revived and novel perspectives on fundamental problems

This Research Topic presents novel and revived perspectives on the fundamental problems underlying psychology's crises in replicability, validity, generalisability and thus, confidence in its findings. Our 15 articles present critical analyses of established theories and practices that are widely used in quantitative psychology and psychological 'measurement'. They show that, contrary to current beliefs, questionable research practices (QRPs) are just surface-level symptoms of much more profound issues that are still hardly discussed.

Uher et al. argue that psychology's crises are rooted in the Questionable Research Fundamentals (QRFs) of many of its theories, concepts, approaches and methods (e.g., of psychometrics)—and therefore cannot be tackled by just remedying Questionable Research Practices (QRPs) as currently believed. The authors emphasise that advancing psychology's theories and philosophies of science is essential for integrating its fragmented empirical database and lines of research. To give new impetus to the current debates, they provide a comprehensive multi-perspectival review of key problems in psychological measurement, highlighting diverse philosophies of science (ontologies, epistemologies and methodologies) that are used in quantitative psychology and pinpointing four major areas of development.

Luchetti explores psychological 'measurement' from a philosophical viewpoint. He notes that, without independent ways for assessing whether a given procedure does, indeed, allow for measuring the intended target property, measurement inherently involves epistemic circularity. From both a modern and a historically-situated perspective, he analyses how Fechner tackled this problem in psychophysics. He shows that Fechner developed a first successful step of a longer-term quantification process. Nevertheless, findings about individuals' sensory perceptions of physical stimuli (e.g., sounds) cannot be generalised to perceptions of all psychical phenomena in lack of evident observable

properties that can be related to the psychical phenomena of interest. The author discusses epistemic circularity as a useful conceptual tool to reflect on the criteria by which measurement standards are regarded as successful in a scientific community.

Kuhbandner and Mayerhofer evaluate limitations of experimental psychology. They critically discuss the field's common assumption that the complexity of the human psyche could be studied in experimentally controlled settings, enabling the identification of law-like behaviours reflective of isolated psychical 'mechanisms'. The authors highlight that even minimal differences in the experimental setup, including differences regarded as irrelevant for a given study, can build up to large unsystematic effects. Moreover, the identification of isolated 'mechanisms', if such were possible, could have no explanatory value given that the psyche functions as a holistic system. They emphasise that the non-mechanistic functioning of higher-order psychical processes cannot be studied experimentally.

Similarly, Mayrhofer et al. interpret the replication crisis primarily as a symptom of an epistemological crisis derived from the mismatch of psychology's quantitative methods with the ontic nature of the psyche. They highlight that failure to replicate findings does not seem to advance the discipline by means of Popperian falsification, yet it also does not bring about Kuhnian paradigm shifts. However, it might address what Lakatos termed the 'hard core' of the discipline's research program. Specifically, the authors argue that over-reliance on quantification in psychology entails a failure to conceptualise its methodological core. A possible solution should aim at a non-quantitative description of psychology's study phenomena that accounts for their observable but unstably quantifiable nature.

In line with this, Linkov, argues that pure ('qualitative') mathematics might be an alternative to measurement. He contends that, in most countries, schools educate students to believe that mathematics equals quantification. Mathematics, however, is the science of abstract structure. Pure mathematics, for example, is the study of mathematical concepts. Its qualitative nature is often turned into quantification and numbers in applied technologies, which can lead to problematic concepts of measurement. Linkov argues that better public understanding of pure mathematics might help the scientific community to distinguish more clearly between qualitative pattern descriptions, quantification and numbers as well as to tackle the ensuing challenges to understanding measurement.

Scharaschkin elaborates similar views in the context of educational assessment. He critically discusses the common conceptualisation of person abilities as latent quantities, as done in many theories of psychological 'measurement' that are aimed at locating a measurand at a point on that numerical continuum. The author suggests that van Fraassen's more expansive view of measurement as location in a logical space provides a more appropriate conceptual framework. Drawing on fuzzy logic and mathematical order theory, Scharaschkin demonstrates a 'qualitative mathematical' theorisation for educational assessments of intersubjectively constructed phenomena (e.g., learner proficiency). This highlights the theory-dependent nature of valid representations of such phenomena, which need not be conceptualised structurally as values of quantities.

Scholz goes a step further and proposes Barad's agential realism as a suitable alternative philosophy of science for quantitative psychology. Contemporary views distinguish between the ontic existence of pre-existing objects of research (entity realism) and the researchers' epistemic approaches for exploring them. The author introduces agential realism, which rejects entity realism and views instead ontic existence and epistemic approaches as entangled and co-created by the researchers. Applied to quantitative psychology, agential realism necessitates the reconceptualisation of common assumptions about 'true scores', context as independent influence factors, the researchers' independence of their objects of research as well as the conception of the research process itself.

Exploring philosophical perspectives on validity, Ramminger and Jacobs discuss the critical role of theory in understanding and evaluating validity in psychological 'measurement'. The authors contrast three positions on validity: Cronbach and Meehl's construct validity, rooted in logical positivism; Borsboom's realist perspective, which highlights causal relationships, as well as Borgstede and Eggert's critique of validity as a concept. The authors contend that, despite their philosophical differences, all three perspectives converge on the essential role of theory-driven approaches in psychological 'measurement'.

Uher provides a comprehensive critique of psychology's overreliance on statistical modelling at the expense of epistemologically grounded measurement processes. She shows that statistics is not measurement because statistics deals with structural relations in data regardless of what these data represent, whereas measurement establishes traceable empirical relations between the phenomena studied and the data representing information about them. Using basic epistemic criteria and methodological principles that underlie physical measurement (e.g., traceability, coordination, calibration), she shows that, in psychological 'measurement' (e.g., psychometrics), many researchers mistake judgements of verbal statements for measurements of the phenomena described and overlook that statistics can neither establish nor analyse a model's relations to the phenomena explored. She elaborates epistemological and methodological fundamentals for establishing genuine analogues of measurement in psychology that consider the peculiarities of its study phenomena (e.g., higher-order complexity, non-ergodicity) as well as those of its language-based methods (e.g., inbuilt semantics).

Arnulf et al. further explore the semantic perspective on the relations between data and study phenomena. They systematically analyse how and why digital language processing can predict psychometric and statistical results fairly accurately even without access to human response data. Reviewing a range of empirical publications that demonstrate this fact, the authors argue that this is because prevalent psychometric analyses capture only the semantic representation of the variables but not the empirical correlates of these variables themselves. The authors highlight that this implies a prevalent category mistake in psychology where 'what can be said' about a phenomenon is mistaken for the phenomenon itself. The ability of technologies, such as large language models, to predict and model response statistics a priori suggests that psychology is

building a semantic rather than a nomological network of variables as commonly assumed.

In their critical analysis of the use of terms in psychology, Hanfstingl et al. emphasise the importance of identifying jingle and jangle fallacies. Jingle fallacies occur when distinct psychological study phenomena are grouped under the same term, whereas jangle fallacies arise when, vice versa, the same study phenomenon is described using different terms. The authors propose a four-step procedure to detect and address issues related to these fallacies, involving problem definition, identification, visualisation and reconceptualisation of the identified fallacies. They highlight that, ultimately, addressing jingle and jangle fallacies requires collective efforts and the incorporation of diverse theories, perspectives and methodologies.

Slaney et al. explore the rhetorical language commonly used in scientific discourse about the theory, validity and practice of psychological 'measurement'. They examine various discourse practices, such as rhetoric (e.g., persuasion), tropes (e.g., perfunctory claims), metaphors and other 'literary' styles as well as ambiguous, confusing or unjustifiable claims. Using conceptual analysis and exploratory grounded theory, they analysed a sample of $N = 39$ articles that were randomly selected from larger article databases representing issues published in 2021 in APA journals across a range of subject categories. The authors identify relevant themes, illustrated with constructive and useful but also misleading and potentially harmful discourse practices.

Using a more classical approach, Reisenzein and Junge introduce a framework to study the intensity of emotions that is based on a realist view of quantities and that combines modern psychometric (latent-variable) approaches with a deductive order of inquiry for testing measurement-theoretical axioms. It relies on Ordinal Difference Scaling (ODS), a non-metric probabilistic indirect scaling technique originally developed to assess sensations, bodily feelings and mental states. The authors discuss the psychological processes involved, including the comparison of stimulus intensities and the role of statistical models in ensuring measurement reliability. The approach bridges theoretical assumptions and empirical methodologies and offers insights for improving the precision of emotion-related assessments.

Brauner, in turn, takes a pragmatic and interesting step away from the necessity to measure purported 'latent constructs'. Instead, he proposes to include several, disparate assessment points in so-called 'micro scenarios' as an integrative contextual method to evaluate mental models and public opinion. He explains how public opinion can be mapped across people and problem spaces, offering practical examples from high-risk technologies (e.g., nuclear power). This approach offers a tool for more informed decision-making, such as in technology development and policy-making.

Paredes and Carré are also concerned with the problems of psychometrics and how these can be remedied. Whereas most approaches focus on statistical and technical best practices for researchers, the authors focus on the challenges that arise from the human-based generation of psychological data. They emphasise the necessity to develop a wider and more nuanced understanding of how different people, communities and cultures interpret and use psychometric 'scales'. Therefore, they propose participatory approaches involving a broader group of stakeholders throughout the measurement process—including researchers, practitioners and the participants themselves.

With our compilation of research papers, we aim to contribute to and stimulate critical debates on quantitative psychology and measurement. We hope that the revived and novel perspectives discussed in these papers will provide good food for thought to motivate and help psychologist to tackle the current challenges and advance psychology as a science.

## Author contributions

JU: Conceptualization, Project administration, Writing – original draft, Writing – review & editing. JA: Writing – original draft. BH: Writing – original draft.

## Acknowledgments

We thank our authors, editors and reviewers for their valuable contributions to this Research Topic. We also thank Alessia Coppola for her outstanding administrative support.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

frontiers | Frontiers in Psychology

TYPE Review
PUBLISHED 21 March 2024
DOI 10.3389/fpsyg.2024.1308098



OPEN ACCESS

EDITED BY
Davide Marocco,
University of Naples Federico II, Italy

REVIEWED BY
Franca Crippa,
University of Milano-Bicocca, Italy
Zhihua Li,
Hunan University of Science and Technology, China

*CORRESPONDENCE
Jan Ketil Arnulf
✉ jan.k.arnulf@bi.no

RECEIVED 05 October 2023
ACCEPTED 27 February 2024
PUBLISHED 21 March 2024

CITATION
Arnulf JK, Olsson UH and Nimon K (2024) Measuring the menu, not the food: "psychometric" data may instead measure "lingometrics" (and miss its greatest potential). *Front. Psychol.* 15:1308098. doi: 10.3389/fpsyg.2024.1308098



# Measuring the menu, not the food: "psychometric" data may instead measure "lingometrics" (and miss its greatest potential)

Jan Ketil Arnulf[1]*, Ulf Henning Olsson[1] and Kim Nimon[2]

[1]BI Norwegian Business School, Oslo, Norway, [2]Department of Human Resource Development, University of Texas at Tyler, Tyler, TX, United States

This is a review of a range of empirical studies that use digital text algorithms to predict and model response patterns from humans to Likert-scale items, using texts only as inputs. The studies show that statistics used in construct validation is predictable on sample and individual levels, that this happens across languages and cultures, and that the relationship between variables are often semantic instead of empirical. That is, the relationships among variables are given a priori and evidently computable as such. We explain this by replacing the idea of "nomological networks" with "semantic networks" to designate computable relationships between abstract concepts. Understanding constructs as nodes in semantic networks makes it clear why psychological research has produced constant average explained variance at 42% since 1956. Together, these findings shed new light on the formidable capability of human minds to operate with fast and intersubjectively similar semantic processing. Our review identifies a categorical error present in much psychological research, measuring representations instead of the purportedly represented. We discuss how this has grave consequences for the empirical truth in research using traditional psychometric methods.

KEYWORDS

**semantic algorithms, semantic networks, nomological networks, latent constructs, natural language processing, measurement, organizational behavior, cross-cultural psychology**

## Introduction

This is a conceptual interpretation and synthesis of empirical studies using semantic algorithms that are capable of predicting psychological research findings *a priori*, in particular survey statistics. The main motive for this study is to sum up findings from a decade of psychological research using text algorithms as tools. As will be shown, outputs from this methodology are now quickly increasing with the advent of powerful and accessible technologies. Available research so far indicates that the phenomenon which Cronbach and Meehl (1955) described as a "nomological network" may, more often than not be of semantic instead of nomological nature. We believe that this confusion has led to decades of categorical mistakes regarding psychological measurement: What has been measured is the systematic representations of abstract propositions in the minds of researchers and subjects, not the purported lawful relationships between independently existing phenomena, i.e., the supposed contents of the construct. Hence the title of this study: measuring the "menu," the semantic

representation, instead of the "food," the subject matter of the representations.

This proposition builds on a set of empirical evidence made possible in recent years through the advancement of natural language processing (NLP) technologies. This evidence will be presented thoroughly in later sections, but we will first give the reader a very brief introduction to the technology and why it matters for social science research. The most famous example of NLP technologies in recent years has been large language models (LLMs) like OpenAI's "ChatGPT" or Google's "Bard." These tools can read inputs in natural language, discuss with human users, and produce texts that are logically coherent to the extent that they can write computer code and analyze philosophical topics.

Users who simply "talk" with the LLMs only meet their human-like responses. They do not have access to the computational workings behind the interface. However, these features are made possible through previous developments in assessing and computing semantic structures in human language. Building on decades of research, NLP approaches have found ways to represent meaning in texts by quantifying linguistic phenomena such as words, sentences and propositions (Dennis et al., 2013). Increasingly, the semantic processing techniques have been found to match or emulate similar processes in humans, narrowing the gap between human and computer capabilities (cfr. Arnulf et al., 2021).

Of particular relevance to the present topic, NLP techniques such as Latent Semantic Analysis (Dumais et al., 1988), Word2Vec (Mikolov et al., 2013), and BERT (Devlin et al., 2018) have been available to measure and compute the semantic structures of research instruments as well as theoretical models and research findings. Without going into details at this point, the mentioned technologies allow us to compute the degree to which variables overlap in meaning (Larsen and Bong, 2016). This has opened a completely new perspective on methodology because it appeared that a vast range of research findings hitherto seen as empirical were instead following from the semantic dependencies between the variables: semantic algorithms can actually predict 80–90% of human response patterns *a priori* based only on the questionnaire texts as inputs, sometimes replicating all information used to validate constructs (Arnulf et al., 2014; Nimon et al., 2016; Gefen and Larsen, 2017; Shuck et al., 2017; Kjell et al., 2019; Rosenbusch et al., 2020; Larsen et al., 2023).

It is important to understand that NLP technologies do not only map and compute wordings of questionnaires, but their calculations also pervade definitions of variables and constructs (Fyffe et al., 2023; Larsen et al., 2023). Since these calculations span the scientific process from empirically collected respondent data to the theoretical argumentation of the researchers, we need to reconsider the distinction between empirical and semantic features of data. The empirical studies to be reviewed here only come about because abstract propositions in the human mind have systematic properties that render them accessible to statistical modelling from text alone. The outline of the present study is as follows:

We will first describe how language processing algorithms can allow *a priori* predictions of response patterns to prevalent, state-of the art measurement instruments in organizational psychology (Arnulf et al., 2018a,b,c,d). Next, we will show how the prediction works across languages and culturally diverse samples (Arnulf and Larsen, 2018). We then use these research findings to re-interpret Cronbach and Meehl's (1955) original concept of "nomological networks" with the more accurate terminology "semantic networks." We argue that many psychological variables do not really "predict" each other in a causal or temporal sense. Instead, they are better understood as re-interpretations of each other as nodes in semantic networks. It is this feature that keeps producing construct identity fallacies (Larsen and Bong, 2016), also called the "jingle/jangle problem."

One peculiar consequence is the empirical demonstration that construct validation conventions tend to lock the explained variance in psychological studies at a constant average of 42% (Smedslund et al., 2022). Another consequence is that semantic networks cannot express empirical truth values (Arnulf et al., 2018a,b,c,d). Semantic networks are prerequisites for the human talent to create arguments and counterfactual hypotheses (Pearl, 2009). This is precisely the reason why we have empirical science, as we need other types of information to falsify hypotheses (Russell, 1918/2007).

Finally, we will point at possible ways to advance from here. Humans display an ability for semantic parsing that is predictable on a level unsurpassed in experimental psychology (Michell, 1994). We posit that statistic modelling of semantic processes is a necessary step to understand that psychological research is itself a revealing psychological phenomenon. The phenomena that will be addressed and discussed in this article are outlined in Figure 1.

## Prediction of empirical statistics *a priori*

Probably the most common approach to empirical psychology is to establish a theoretical relationship between two or more defined constructs, operationalize the constructs as variables and collect some types of data (Bagozzi and Edwards, 1998; Nunally and Bernstein, 2007; Borsboom, 2008; Bagozzi, 2011; Michell, 2013; Vessonen, 2019; Uher, 2021b). The testing of the hypotheses, and hence the theories, hinges in the measurement data fitting the predictions, that in turn belong to the argued theories (Popper and Miller, 1983; Jöreskog, 1993). The purpose is to allow a quantitative description of the relationship between the variables, based on the numbers obtained as measurements.

Following predominant philosophy of science, the assumption is that reasonably argued theoretical relationships should withstand attempts to falsify them (Popper, 1959). The falsification could take two steps: First, a statistical rejection of null hypotheses showing that the numbers are reasonably non-coincidental, and secondarily the hypotheses are supported (by not being disconfirmed).

Hence, psychological research abounds with complex and elaborate statistical models that either stepwise or in one sweep take all these concerns into consideration (Lamiell, 2013). If the numbers fit the statisticians' model requirements, the findings are generally accepted as "empirically supported." What this should imply, is that the measurement results came about as independent observations from the theoretical propositions.

A number of research traditions have over the years voiced doubt about this independence. The doubt has largely taken two forms. The first type of doubt in the data independence came from criticism of the widespread use of quantitative self-report instruments. Starting already upon Rensis Likert's adoption of quantitative response categories to questionnaires in 1932, other researchers were concerned

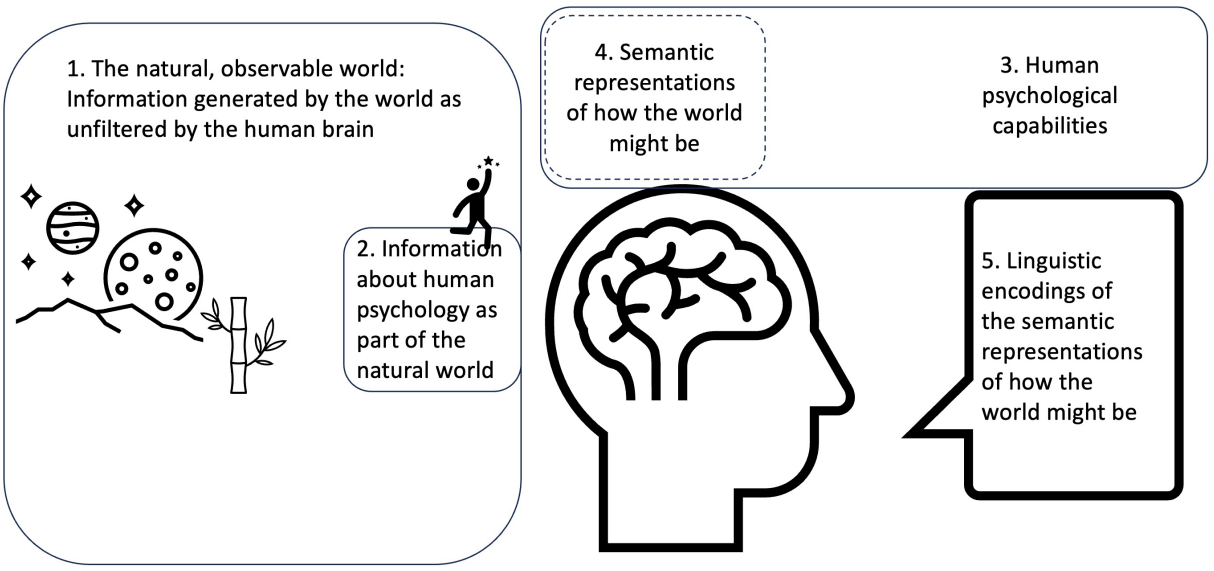



FIGURE 1
Empirical science should ideally tap into the features of the world that we cannot see (1). Psychology takes aim at a piece of this (2), but often stops at the doorstep by collecting information about how we represent the world (4). This is itself a product of our psychological apparatus (3) and we need it to talk about what we see and find (5), but it is not itself empirical certainty about 1 and 2.

TABLE 1 Semantic similarities between four statements about gardening.

| | It is raining | It is snowing | The sun is shining | The lawn is wet |
|---|---|---|---|---|
| It is raining | 1 | | | |
| It is snowing | 0.93 | 1 | | |
| The sun is shining | 0.43 | 0.41 | 1 | |
| The lawn is wet | **0.80** | **0.73** | **0.44** | 1 |

Relationships with dependent variable in bold.

about the nature of the ensuing numbers as well as about the value of self-reported responses across many domains of inquiry (LaPiere, 1934; Drasgow et al., 2015).

The second type of doubt has targeted a broader and more conceptual side to psychological research, regardless of the method applied. What if the empirical data collection is set up to replicate something that is necessarily true? Many such situations are conceivable, such as finding out whether people who experience something unexpected will turn out to be surprised (Semin, 1989; Smedslund, 1995). While some such examples may be blatantly obvious, incisive theoretical analyses have found several instances of more indirect versions of this where the dependent and independent variables are found to be parts of each other's definitions (van Knippenberg and Sitkin, 2013). Where variables are conceptually overlapping, they will also be statistically related if measured independently.

Both of these concerns allow us to state a very precise prediction: When research instruments or designs ask questions where the answers are given by the meaning of the questions used, the resulting statistics should be explainable by the texts. More precisely, the information contained in the definitions of constructs, variables and research questions comes back as the observed statistics (Landauer, 2007). When this happens, the measures are not independent information that fit the theories. The measures are measuring the semantic properties of theoretical statements in a self-perpetuating loop.

A simple example may illustrate how semantic algorithms can model empirical data. Assume that we are asking respondents about the condition of their lawn, rating the item: "The lawn is wet." To "predict" this variable, we ask people to rate three other variables: (1) "It is raining," (2) "It is snowing," and (3) "the sun is shining." We can run this example with LSA at the openly available website http://wordvec.colorado.edu/, and the results are displayed in Table 1. A mere semantic analysis of the statements is predicting the likely outcome of this empirical exercise: If it rains, the lawn is likely to be wet. By snow it is almost as likely, but if the sun shines, it is less likely to be wet.

The important point here is not the absolute values, but the mutual quantitative relationships between variables. These semantic values can be compared to correlations or covariances, but they are not meaningful as single data points, only as relationships. The results in Table 1 are blatantly obvious but the same principles hold across far less obvious data structures.

At the moment of writing, studies demonstrating semantically predictable research findings and picking up at an increasing pace covering state-of-the-art research instruments in leadership and motivation (Arnulf et al., 2014), engagement, job-satisfaction and well-being (Nimon et al., 2016), the technology acceptance model (Gefen and Larsen, 2017), job analysis (Kobayashi et al., 2018), personality scale construction (Abdurahman et al., 2023; Fyffe et al., 2023), entrepreneurship (Freiberg and Matz, 2023) personality and

mental health (Kjell K. et al., 2021; Kjell O. et al., 2021) or even near-death-experiences (Lange et al., 2015). Overlapping meanings between a vast group of constructs have been demonstrated (Larsen and Bong, 2016) and new scales can be checked for overlaps (Rosenbusch et al., 2020; Nimon, 2021).

Some of these studies will be explained in more detail below, but we first need to recapitulate some of the features of latent constructs that allow such predictions from the measurement texts alone.

## The latent construct and its cognitive counterpart

Up until the mid-1950s, mainstream psychological research was dominated by a behaviorist and/or positivist view on what constituted legitimate empirical variables (Hergenhahn, 2009). Invisible, inferred psychological phenomena like thinking and emotions were regarded with theoretical suspicion as they could not be observed directly. This changed considerably with the "cognitive revolution" that in many ways paralleled the growing understanding of information and communication theories (Shannon and Weaver, 1949; Pierce, 1980). Borrowing from the idea of "operationalism" in physics, psychology gradually warmed up to the idea of studying phenomena inside the organism by adopting "hidden" variables or through lines of argumentation that would result in "constructed" variables (Bridgman, 1927; Boring, 1945). One milestone came with Cronbach & Meehl's paper on the statistical criteria for "construct validation" (Cronbach and Meehl, 1955). This contribution was to become the cornerstone of APA's test manual guideline for construct validation, as the latent variable became an established feature of empirical psychology (APA, 2009; Slaney, 2017b).

Acceptance in mainstream methodology notwithstanding, latent variables still have the peculiar feature that they cannot be observed. They will always have to be inferred from operationalizations, i.e., other more empirically accessible observations that point towards the existence of a common factor. Moreover, their ontological status has never been settled within the psychological sciences (Lovasz and Slaney, 2013; Slaney, 2017a). With the advent of desktop computing in the 1980s, factor analysis became a tool for everyone and methods for modelling these proliferated (Andrich, 1996). The proliferation of statistical methods brought about a similar proliferation of new latent constructs (Lamiell, 2013; Larsen et al., 2013). Such rapid increase in constructs raised another hundred year old problem in psychological theorizing (Thorndike, 1904): How and when do we know if two theoretical variables are the same, even if they carry the same name? Or how can we know that two groups of researchers are really working on the same problem, simply by knowing the name of the construct they are working on?

This question has been named the "construct identity problem" and points to a problematic but interesting feature of human cognition that also affects researchers (Larsen and Bong, 2016): What's in a name? It is obviously possible for humans to believe that two statements are distinct, even though they are making the same point. The all-too-human confusion on this point is a major feature in the research on decision making such as the seminal research of Kahneman and Tversky on framing (e.g., Tversky and Kahneman, 1974; Kahneman and Tversky, 1982). Recent research indicates that such problems, often referred to as the "jingle-jangle"-problem, are very real phenomena in psychological research (Nimon et al., 2016). While digital text algorithms can detect and differentiate construct identity fallacies across large swaths of constructs, humans have in fact a hard time detecting such similarities (Larsen and Bong, 2016).

Thus, the latent and elusive nature of constructs go together with a cognitive handicap in humans, the fact that we are often oblivious about overlaps and relationships between the constructs. This renders psychology and related disciplines vulnerable to linguistic fallacies since most latent variables shaping research are also everyday concepts that are known and taken for granted by most people (Smedslund, 1994, 1995). Psychological research is concerned with learning, thinking, emotions, perception and (mostly) easily understandable constructs in healthy and disturbed personalities (Haeffel, 2022). But can we be certain that everyday concepts can be treated as fundamental entities of psychological theory – latent variables – just because their measurement statistics correspond to APA requirements from 1955 (APA, 2009)?

Or is it time to move on, to see that we have been doing research on questions that were largely determined – and in fact answered – by our own cognitive machinery? What would psychology look like if it could peek beyond the "manifest image" of the latent constructs and the computational machinery that makes us construct them (Dennett, 2013; Dennett, 2018)? We will now turn to discuss the empirical findings that could help us find such a perspective.

## Predicting leadership constructs

By 2014, the world's most frequently used questionnaires on leadership was the Multifactor Leadership Questionnaire (MLQ, Avolio et al., 1995), figuring in more than 16,000 hits on Google Scholar. A study published that year (Arnulf et al., 2014) showed that the major parts of factor structures and construct relationships in the MLQ was predictable through text algorithms, using only the item texts as inputs. By running all the questions (or items) of the questionnaire through Latent Semantic Analysis (LSA) (Dennis et al., 2013) similar to the procedure in Table 1, it was possible to calculate the overlap in meaning among all items involved. The LSA output matched the observed correlations almost perfectly. Depending on the assumptions in the mathematical models, it was possible predict around 80 to 90% of the response patterns of humans from semantic similarities (Arnulf et al., 2014).

## Individual response patterns

Given the possibility that sample characteristics are predictable *a priori*, does this also apply to individual response patterns? Semantic predictions cannot know which score level a given respondent will choose when starting to fill out the survey. But, since all items are linked in various ways to all other items (weakly or strongly), it should theoretically be possible to infer something about subsequent responses after reading a few initial ones? Another study addressing precisely this question discovered that knowing the first 4–5 items of the MLQ allowed a fairly precise prediction of the 40 next responses (Arnulf et al., 2018b). In other words, the semantic relationships are not restricted to samples, they emerge already as features of individual

responses. This amazing semantic precision was already predicted by unfolding theory in the 1960s (Coombs and Kao, 1960).

## Human linguistic predictability

Reading and parsing sentences comes so easily to people that it feels like reacting directly to reality. And yet, tasks like reading, comprehension, and responding to survey items are all behavioral processes based on psycholinguistic mechanisms in the brain (e.g., Poeppel et al., 2012; Krakauer et al., 2017; Proix et al., 2021). The first central feature of the semantic processing is remarkable but easily overlooked: It provides a rule-oriented predictability to people's verbal behavior unlike any other behavior systems known in psychology except biological features of the nervous system, allowing humans to easily parse and rank texts like survey items along their semantic differentials (Michell, 1994).

Therefore, a semantic representation of Likert-scale survey items may allow us to predict the statistical patterns from both samples and individuals. Since these levels of predictability exceed most other processes in psychology (Michell, 1994; Smedslund et al., 2022), it is highly likely that semantic similarity numbers are matching and quite probably mirroring the outputs of the linguistic processes of the brain itself (cfr. Landauer, 2007). However, the process must take place on the semantic levels, not the basic linguistic parsing. The cognitive features of constructs seem relatively independent of the words used to encode them. We will show this by showing how semantic algorithms can model constructs across cultures and languages.

## The cultural invariance of semantic relationships

The study on semantic features of the MLQ described above (Arnulf et al., 2014) had an interesting design feature: The algorithm predicting the numbers worked on English language items as inputs and was situated in Boulder, Colorado while the respondents filling out the survey were Norwegians, filling out a Norwegian version of the MLQ. The algorithm knew nothing about Norwegian language or respondents. While previous research had established that LSA could work across languages (Deerwester et al., 1990), it was not obvious that propositional structures in research topics such as leadership would be statistically similar across linguistic lines. It turns out to be possible to demonstrate this similarity across even greater divides.

One study was designed to demonstrate how propositions about leadership appear as universally constant across some of the biggest linguistic and cultural divides that exist (Arnulf and Larsen, 2020). The method was applied to a very diverse group of respondents from China, Pakistan, India, Germany, Norway, and native English speakers from various parts of the world. The non-English speakers were divided into two equal groups, one responding in their mother tongue, the other half responding in English. Again, the semantics were calculated with LSA, using English language items only.

For practical purposes, the LSA output performed completely unperturbed by the linguistic differences. As in the first study, the LSA numbers almost perfectly predicted the response patterns of all groups that responded to items in English. The groups responding in other languages were slightly less well predicted and one might have speculated that there were "cultural" differences after all. However, the methodological design allowed comparisons of all groups responding in either English or their mother languages and they did not share any unique variation. In other words, there were no commonalities attributable to culture or other group characteristics. The differences in predictability across these experimental groups could only be explained by imprecise translations of the items. The propositional structures of the original instrument were picked up and used uniformly across all respondents. In other words, the propositions can be modelled statistically independently of the language used to encode them.

Both algorithms and human respondents reproduce the relationships of abstract meaning among the items. In that sense, the patterns are abstract transitive representations of the sort that "If you agree to A, you should also agree to B, but disagree to C…" as predicted by unfolding theory (Coombs and Kao, 1960; Coombs, 1964; Michell, 1994; Kyngdon, 2006). The culturally invariant feature of semantics is very important because it shows that semantic networks are prerequisite for language, but not language itself. The system of propositions hold in any language, including sign language (Poeppel et al., 2012). This is the whole point of accurate translations and back-translations in cross-cultural use of measurement scales (Behr et al., 2016). We posit that the data matrices from humans in any language match that of the LSA algorithm not because any language is correct, but *because their deeper semantic structure have identical mathematical properties* (Landauer and Littman, 1990; Landauer and Dumais, 1997; Landauer, 2007). In turn, this raises another problem because the constructs are then never truly independent – they derive their meanings from their mutual positions in the semantic network, as we will show next.

## The not so empirical variables

While "causality" is a strong word in the sciences (Pearl, 1998, 2009), most study designs in quantitative social science explore how one variable changes with changes in the another. To study an empirical relationship is usually taken to mean that the focus variables are free to vary, and that quantitative regularities between the two were unknown or at least uncertain prior to the investigation.

Explorations of the semantic relationships between variables indicate that frequently, the variables involved in psychological studies are *not* independent of each other. To the contrary, they may actually be semantic parts of each other's definitions and belong to the same phenomenon (Semin, 1989; Smedslund, 1994, 1995; van Knippenberg and Sitkin, 2013; Arnulf et al., 2018c). To underscore this point, the above mentioned study of leadership scales found cases where the semantic algorithms predicted the relationships among *all* the involved variables (Arnulf et al., 2014). The predicted relationships were not restricted to the MLQ but spilled over into all other variables argued to be theoretically related to transformational leadership. Semantic patterns detected the relationships between independent variables (in this case, types of leadership), mediating variables (in this case, types of motivation) and dependent variables (in this case, work outcomes).

When this happens, constructs are in no way independent of each other. They are simply various ways to phrase statements about working conditions that overlap in meaning – a sort of second-order

jingle/jangle relationships (Larsen and Bong, 2016; Nimon, 2021). In our example, it means that definitions and operationalizations of work imply a little bit of motivation. The definition of motivation, in its turn, implies a little bit of work effort. But the definitions of leadership and work effort show less overlap. The statistical modelling makes semantic relationships *look* like empirical relationships, where leadership seems to affect motivation, in turn affecting work outcomes. But this is just the way we talk about these phenomena, just as Thursdays need to turn into Fridays to ultimately become weekends.

## Lines of reasoning – nomological or semantic networks?

Our failure to distinguish between semantic and empirical relationships is itself an interesting and fascinating psychological phenomenon. When we are faced with a line of reasoning, it may seem intuitively appealing to us. Our need for empirical testing stems precisely from the fact that not everything that is arguable is also true as a fact. Counterfactual thinking is crucial to human reasoning (Pearl, 2009; Pearl and Mackenzie, 2018; Mercier and Sperber, 2019). But, conversely, some of the facts we find are probably true simply because they are arguable – they are related in semantic networks (cfr. Lovasz and Slaney, 2013).

This crucial point was raised by Cronbach and Meehl in their seminal paper that founded the psychometric tradition of construct validation (Cronbach and Meehl, 1955). We will show how the semantic properties of constructs can be explained by a re-interpretation of Cronbach and Meehl's "nomological network," spelled out as six principles and explained over two pages (Cronbach and Meehl, 1955, pp. 290–291).

A bit abbreviated, the six principles state that: (1) A construct is defined by "the laws in which it occurs." (2) These laws relate observable quantities and theoretical constructs to each other in statistical or deterministic ways. (3) The laws must involve observables and permit predictions about events. (4) Scientific progress, or "learning more about" a construct consists of elaborating its nomological relationships, or of increasing its definite properties. (5) Theory building improves when adding a construct or a relation either generates new empirical observations or if it creates parsimony by reducing the necessary number of nomological components. (6) Different measurement operations "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable.

We propose that these six principles do not spell out an empirical nomological network. The word "nomological" as invented by Cronbach and Meehl means "governed by laws" (from the Greek word "nomos" meaning law) and would imply that there are lawful regularities between the constructs. However, causal laws are never described between psychological constructs – they are always modelled as correlations or co-variances. In fact, at the time the "nomological networks" were proposed, psychological statistics was ideologically opposed to laws and causation under the influence of Karl Pearson (Pearson, 1895, 1897; Pearl and Mackenzie, 2018). Instead, the six principles outlined by Cronbach and Meehl perfectly describe the properties of a semantic network, where all nodes in the network are determined by their relationships to each other (Borge-Holthoefer and Arenas, 2010).

Here follows our re-interpretation of Cronbach & Meehl's six principles as semantic networks (principles annotated with an "S" for semantics):

(S1) A construct is defined by "the semantic relationships that define it." (S2) These semantic relationships explicate how the construct is expressed in language, and how it may be explained by other statements. (S3) The relationships must involve concrete instances of the constructs linking them to observable phenomena. (S4) Scientific progress, or "learning more about" a construct consists of expanding its semantic relationships, or of detailing its various meanings. (S5) Theory building improves when adding another construct can be argued to expand the use of the construct, or if it creates parsimony by reducing the number of words that we need to discuss it. (S6) Different measurement operations "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable.

Next, we will show how the semantic network works in theory and research on the construct "leadership," using a commonly used definition provided by Northouse (2021, p. 5):

(E1) "Leadership is a process whereby an individual influences a group of individuals to achieve a common goal." (E2) This implies a series of interactions between one human being and a group of others, where the first individual has a wide range of possible ways to influence cooperation in the group of others. (E3) There must be some form of communication between the leading individual and the others, as well as involving a time dimension that allows a process to take place. (E4) Scientific progress may consist of explicating what "influence" may mean, and also about what a "process" might be. (E5) Theory building improves if other concepts such as "motivation" can expand the use of the construct, or if it creates parsimony by explaining what the group will be doing instead of having to describe the behaviors of all group members in detail. (E6) The definition covers agency in the form of influence, groups of individuals, time lapse (process) and end states (goals). Other ways of describing the process may overlap if capturing agency, influence, groups, time laps and end states in different ways.

The two ideas of nomological vs. semantic networks are strikingly similar but have very different implications. Semantic networks do not require any other "laws" than precisely a quantitative estimate of overlap in meaning. In parallel, prevalent techniques for construct validation never require data that go beyond correlations or covariations (Mac Kenzie et al., 2011). Our main proposition is that if semantic structures, obtainable *a priori*, allow predictions of the observed relationships between variables, then the network properties are probably rather semantic than nomological.

This distinction between nomological and semantic networks is crucial to understand the true power of semantic algorithmic calculations. The semantic networks between concepts (or, for that matter, "constructs") are what allows us to reason and argue (Mercier and Sperber, 2019). It is precisely this feature of concepts that make up the logical argumentation in the "theory" part of our scientific productions. Moreover, the very idea about "constructs" that Cronbach developed was taken from Bertrand Russell's argument that one should be able to treat phenomena as real and subject them to science if they can be inferred logically from other propositions (Slaney and Racine, 2013).

At the same time, the fact that there exists a semantic network that we ourselves easily mistake for a nomological one reveals a very intriguing feature of human psychology: Our abstract, conceptual thinking is following a mathematically determined pattern, but

we ourselves are not aware of it (Dennett, 2012). The semantic algorithms allow us to model structures of abstract propositions simply from the properties of sentences – which is most likely what the human brain itself is doing (Landauer and Dumais, 1997; Landauer, 2007).

Oddly, what this implies is that much of the assumedly empirical research implying construct validation is not actually empirical (Smedslund, 1995, 2012). Instead, this type of research is more akin to a group-sourcing theoretical endeavor, establishing what can reasonably be said about constructs in a defined population of respondents. Hence the title of this article: The measurements collected are not measuring what the scales are "about." Instead, the measurements are measuring what we are *saying* about these things. It is a mistake of categories, mistaking the menu for the food or the map for the terrain (Russell, 1919; Ryle, 1937).

The good thing about this type of research though is that it can be seen as a theoretical exercise, establishing that the theoretical relationships between items make sense to most people. For example, the cross-cultural research on transformational leadership that claims to find similar factor structures across cultures tells us only that all people agree how statements about leadership hang together. This is far from establishing contact with behavior on the ground, but it is precisely the essence of a theoretical statement.

If constructs, or the semantic concepts that make them up, have reasonably stable properties that define them, then there must be deterministic procedures in the brain to evaluate the overlap in meaning of statements. Whether software or wetware arrive at these evaluations in the same way is not important. The important part is that the coherence of statements is a mathematically representable structure, like Landauer has shown for LSA (2007) or Shannon has shown to be the case for general information systems (Shannon and Weaver, 1949; Pierce, 1980).

Due to this, it is possible to mutate constructs into propositions that overlap in meaning despite being encoded in different words. In the early years of analytical philosophy, the German logician Gottlob Frege was able to show this (Frege, 1918; Blanchette, 2012). He made a crucial distinction between "reference" and "meaning," a precursor to the jingle/jangle-conundrum: Different sentences may refer to the same propositional facts even if they have no words in common. Their meaning however can be slightly different, capturing many layers of linguistic complexity.

In this way, human subjects often miss how data collection designs simply replicate the calculations in the semantic network. What we see here is the constructive feature of our semantic networks: They allow systematic permutations of all statements that can be turned into latent constructs precisely because they have systematic features.

## The 42% solution

Interestingly, it turns out that when constructs share less than 42% of their meaning, humans experience them semantically separate, and therefore possible targets of empirical research. Evidence for this has been found in a study that analyzed a wide range of constructs and across psychological research (Smedslund et al., 2022).

This study reviewed all the publications listed in the PsycLit database (and that referred to explained variation in the abstract) and found that the average explained variation every year since 1956 was exactly 42%. This number kept being remarkably constant throughout seven decades and 1,565 studies, including both self-report and independently observed data. By reconstructing 50 of them with only semantic data, it became evident why the number has to approximate 42. This is simply the average percentage of semantic overlap between any construct and its neighboring constructs in studies where the constructs are separated by factor analysis. In this way, psychological method conventions have built a scaffolding around our conscious experience of semantic similarities.

One may think of the mechanism this way:

Variables – or measurement items – with obvious overlap in meaning will always be grouped under the same construct. So, in factor analysis, such highly overlapping clusters will emerge as single factors. In the same way, other factors that are included in the analysis will have to appear with much smaller cross-loadings. Different schools of methodology have different cutoffs here, but usual benchmarks are minimum 0.70 for within-factor loadings and maximum 0.30 for cross-loadings. By this type of convention, most studies will publish ratios of within-and cross-loadings around these values. If one divides cross-loadings of 0.30 with within-factor loadings of 0.70, one gets exactly 42%. In plain words, the cross-loadings consist of semantic spillovers from each construct into its neighbors, allowing on average 42% shared meaning between constructs.

The reconstruction of 50 such studies using purely semantic processing of items and/or construct definitions made it clear that the semantic structure alone will yield a mutual explained variance of around 42%. This is another indicator that Cronbach and Meehl's network is not "nomological" but most probably semantic.

However, the most important aspect of this discovery is the implication for psychological theory and epistemology in that the explained variance between constructs is locked within two other features of semantic processing: If two variables are too semantically distant, they will rarely be of interest (cannot be argued to have a relationship), and so they will probably not be researched. Conversely, if their representations have too tight semantic connections, they will be perceived to be the same construct or at best facets of the same.

In this way, the structure of semantic relationships will prepare researchers in the social sciences to design studies in a range from a few percent to maximum 42% overlap, which is what we find to be the average case across all studies reporting percentages of explained variance in the abstract or key words. We must assume that relationships of less than 42% overlap are not immediately obvious to humans as being systematically related through semantics – but they still are.

But why should we care about the difference between a semantically constructed entity and an empirical discovery, if both discoveries seem illuminating and true to humans anyway? The answer is actually alarming – semantic networks do not care whether a calculated relationship is "true" or not. It only maps how propositions are mutually related in language. Therefore, it is definitely possible to propose falsehoods even if the propositions make sense to speakers, a key condition for undertaking scientific investigations (Russell, 1922; Wittgenstein, 1922; Pearl, 2009).

## How semantic networks are oblivious to truth values

To understand this, it is useful to think about theoretical variables from two different perspectives. From one perspective, we are

interested in what variables are about (Cohen et al., 2013). Researchers may be interested in how much we like our jobs, if we are being treated fairly by our employers or whether we think politicians should spend more on schools. When humans respond during data collection, their responses are about what the questions are about (Uher, 2021b).

But from a different perspective, the researchers are often only interested in whether two or more variables are related, no matter their actual strength or value (McGrane and Maul Gevirtz, 2019). This type of relationship is what correlations and covariances are built on and are most often the focus of psychological research. Note that such numbers are only quantifying the relationship itself, but abstracted from what the variables were "about" (Lamiell, 2013; Uher, 2021a).

This problem is most prevalent in research relying on verbal surveys such as Likert-scales (but not restricted to them). Consider two different persons, having different opinions on two questions. One person is giving off the enthusiastic responses, maybe ticking off 6 and 7 on two questions. The next respondent is negative and ticks off only 1 and 2 on the same two questions. From the point of view of their attitude strength, these two persons are clearly different and on opposite ends of the scale. But their systematic relationships with the two questions are exactly identical and they will contribute to the same group statistics and the same correlations in exactly the same way. This is of course the essence of correlations and should not matter if the numbers keep their relationships with their "measurable" substrates (Mari et al., 2017; Uher, 2021a). Looking at measurement from a semantic point of view, this does not seem to happen:

One study using semantic algorithms found a way to tease apart the attitude strength in individual human responses - what a variable is "about" – and the pure relationships between the variables, regardless of their contents (Arnulf et al., 2018a,b,c,d). By differing between the information about how strongly people feel about something, and the mere distance between the variables, it seemed that only the distances between scores played a role in the statistical modeling, not the absolute score levels. Most importantly, this is the part of statistical information that relates most strongly to the semantic structure of these variables. This implies that the way we model propositions semantically is independent from believing that they are true. In fact, commonly used statistical models seem to leave no information left about the topic respondents thought they were responding to. What the models contain are the mutual representations of the variables as propositions. This can be no coincidence as these structures probably mirror their mathematical representations in our cognitive apparatus.

One of the most ingenious yet least understood features of human cognitive capabilities is how we can think, formulate and communicate a near-to infinite range of propositions (Russell, 1922; Wittgenstein, 1922; Wittgenstein, 1953). The possibility to pose hypothetical, competing and counterfactual propositions is probably the very core of causality as understood by humans (Pearl, 2009; Harari, 2015). A crucial condition for "strong artificial intelligence" is arguably the implementation of computational counter-factual representations (Pearl, 2009; Pearl and Mackenzie, 2018).

Thus, while much of the semantic structures uncovered by psychological science might not tell us much about the outside world – what the constructs are "about" (the references), this discovery might actually open up another very interesting perspective. Semantic modelling may help us understand the human mind, and in particular that of the scientists themselves.

## Why algorithms perform as a one-man-band social scientist

Two recent conference papers (Pillet et al., 2022; Larsen et al., 2023) have explored this along two steps: The first step hypothesized and found that the semantic patterns can be used to determine correct operationalizations of constructs. By applying a layer of machine learning on top of the LSA procedures the algorithms could predict correctly which items belong to which construct in a sample of 858 construct-item pairs.

The next step was a test of how the algorithms do compared to humans in the item-sorting task recommended in construct validation, determining which items would make the best fit with theoretically defined constructs (Hinkin, 1998; Hinkin and Tracey, 1999; Colquitt et al., 2019). The algorithms seemed to perform as well as the average humans in deciding if items belong to constructs or not.

If we compare this with the performance in the previously cited articles, we can see that the language algorithms are able to predict data patterns that range from construct definition levels via item correspondent levels (Larsen et al., 2013; Rosenbusch et al., 2020; Nimon, 2021), down to patterns in observed statistics bearing on construct relationships and correlation patterns from human respondents (Arnulf et al., 2014; Nimon et al., 2016; Gefen and Larsen, 2017; Arnulf et al., 2018a,b,c,d).

To sum it up, the semantic algorithms seem able to predict theoretical belongingness of items, the content validity of the items, and the factor structures emerging when the scales are administered. The algorithms can predict individual responses given a few initial inputs, as well as the relationships among the latent constructs across the study design. Taken together, the algorithms seem to trace the systematic statistical representation of the whole research process – from theory to measurements, and from measured observations to variable relationships and factor loadings.

This indicates that there must exist a main matrix within which all the other definitions and measurement issues take place. The whole research process is embedded in semantic relationships from broad theoretical definitions and relationships, through the piloting efforts in sampling suitable items all the way to the final emergence of factor loadings.

This semantic matrix is the very condition for humans to communicate in language. For a word to be a meaningful concept, it needs to be explainable through other words. There is no such thing as a word in isolation. Thus, the phrase "you shall know a word by the company it keeps" actually works in the opposite direction: Words derive their meanings from being positioned relative to their neighbors (Firth, 1957; Brunila and LaViolette, 2022) in the semantic matrix of humans. All latent constructs are embedded in a calculable network which needs to have stable representations across speaking subjects.

At first glance, the requirement of stable semantic networks seems to contradict the differences between people involved in the process of generating measures and those responding to them. There are highly specialized researchers, there are purpose-sampled piloting groups in the development phase and there are the final targeted groups of respondents. There are even controversies in the literature as to whether the test samples in the piloting phase should be experts or lay people.

The semantic network does not seem to be disturbed by this in a major way. It appears so rigidly identical across humans that it feels like a manifestation of nature itself. How inter-subjectively constant is the semantic network really, and can it be computationally addressed?

## Semantic networks across respondents

The methodological gold standard of construct validation in psychology has arguably been the paper of Campbell and Fiske in 1959, claiming that only multi-trait, multi-method (MTMM) designs can estimate measurement errors to an extent that allows the true nature of a construct to be modelled across measurements (Campbell and Fiske, 1959). This is the traditional core of construct validation. By measuring a phenomenon from several angles (often referred to as "traits") and using several methods or sources of information (referred to as "methods") – we can see if a phenomenon has an existence relatively independent of the ways we measure it (Bagozzi and Edwards, 1998; Bagozzi, 2011).

A study building on five different datasets and involving constructs from four different leadership theories, investigated how semantic relationships appear can be modelled within a traditional MTMM framework (Martinsen et al., 2017). In one variation of this design, three different sources rated the appearance of leadership along three different facets or traits of leadership. The semantic representations of the items (generated in LSA) were added to the modelling procedure. In practice, this implied that a manager was rated by him-or herself, by a higher-level manager, and by a subordinate. The design was replicated five times with different people and different constructs. For all datasets, the semantic properties of the relationships between the measurement items were added to the model. The purpose was to establish whether the semantic representations were trait or methods effects, or if they simply captured the errors.

The numbers calculated by semantic algorithms were, in a first step, significantly correlated with the empirical covariance matrix. After fitting the MTMM model, the model implied matrix and its three components were still correlated with the semantic measures of association on a superficial level in four of the five cases.

As the analysis split the covariance components into source, method, and error, the semantic values were present in the trait components in three out of four studies but with only negligible traces in the methods components. The semantic predictability of response patterns was most clearly found in the trait components, or in other words: The validity of a latent construct is equal to its semantic representation – across all the respondents. The semantic properties *are* the construct, they are not an approximation of it.

This became distinctly clear by computing a completely new model of the data, called the restricted-error-correlation named REC-MTMM (Satorra et al., 2023). This model had a near-to perfect fit with the data. The REC-MTMM model implies that the observed sample covariances decompose as the sum four covariance elements of trait (T), source (S), REC parameter (REC), and residual (R) components. The model is accurate enough so that the residual does not contain relevant information.

*We believe that the REC-MTMM correlations and parameters are the imprints of semantic associations.* The fact that parameters can be restricted to be equal across respondents indicates that the respondents are remarkably synchronized in their way of reading the items. This holds even as they *rate* the items differently. In fact, respondents in the three different sources were only partially in agreement about the level of leadership exercised by the person they rated – their attitude strength.

Where their agreement was beyond any doubt, was in being unified in their semantic coherence with the trait characteristics. They all agreed that the items, mutually, had the same meanings. *What this implies, is that no matter how diverse the respondents' experiences of their situations was, they would always unite in a linguistic behavior describing a possible situation.* They were in fact endorsing the properties of the semantic network, just not agreeing about the truth value of what the items proposed.

To understand the implication of this, consider a law case brought up before a court. Someone is accused of theft. All involved – the prosecutor, the defendant, and the judge – will agree that the law categories of robbery, theft or innocence exist and what they mean, but will disagree whether they actually apply in this particular case. In the same way, the respondents of our studies agreed about the various possible categories of leadership but would not always agree of the rated person was "guilty" of this type of leadership.

The REC-MTMM model is effective in bringing out the inter-subjective nature of the semantic network as a common interpretational framework for all people implied. When the fit statistics of the model are as impressive as we find here, it means that we have captured the data generating process itself. It is semantic modeling of the construct and its representation in the language of the respondents that drives this process. The salient function of the network is to provide a common conceptual framework from which speakers can communicate their assessments. Note however in line with what has been described above that the network calculations are indifferent to the truth values of the subjects as long as they can describe the situation in terms of the involved concepts or constructs.

This is the final feature of the semantic network that we want to list in this discussion of the phenomenon. It is a fundamental, intersubjective function with a predictive capability that is beyond anything else in psychology. The involved respondents have all sorts of opinions about a prevailing situation, but they all seem to agree that the situation can reliably and validly be discussed in terms of the involved constructs, as expressed in the REC-MTMM model. A model is a theory of the processes that gave rise to the data that we see, and the REC model taking semantics into consideration makes exhaustive use of all information available.

## Peeking past the semantic matrix: empirical questions

As the almost omnipresent influence of the semantic network is laid bare, one could easily wonder if almost all psychological phenomena can be predictable *a priori* through semantic relationships. What is there left to detect in terms of empirical questions?

This question is maybe one of the most pressing challenges to overcome for psychology and many other social sciences to move forward. If we keep on conducting research on relationships that are already embedded within the semantic network, we will be "addicted to constructs" (Larsen et al., 2013) forever. This practice is very similar to publishing each entry in the multiplication tables. As most children understand around the age of ten, you cannot "discover" the entries of the multiplication table – discovery is superfluous if the numbers are given by applying multiplication rules on the symbols.

One major psychological research question is to explore and describe the nature of the possible neurobiological foundations and

its impact on how we represent the world. The ability of the human linguistic system to detect and encode abstract information could arguable be one of the brain's most advanced features. The profound grasp of semantic representations that can be evoked and processed in most normal people is so precise and automated that we mostly take it for granted (Poeppel et al., 2012). The features of the semantic matrix are probably experienced like the nature of numbers, where humans have struggled for ages to determine whether the numbers are a part of nature or a human invention. In fact, most people probably have the feeling of being in direct contact with reality when coming in contact with the precise and solid patterns provided by semantics.

Yet, the semantic patterning of abstract propositions is no more a feature of nature than the longitudes and latitudes of geography. This delicate intertwining of our semantic representational system with the way that we describe and discuss the world is precisely the reason why it is so hard to discuss and grasp in our scientific findings (Russell, 1922; Wittgenstein, 1922; Mercier and Sperber, 2011). In this sense, semantic representations are to abstract thinking what Dennett (2013) has called the "manifest image" of the physical world. It is nature's remarkably engineered cognitive illusion demanding its own empirical research field.

Another important development would be to start using the semantically calculable relations as the starting point of our scientific investigations. If psychology is to "stop winning" (Haeffel, 2022), and move towards non-obvious expansion of our knowledge, we must stop being impressed by discovering relationships that are knowable *a priori* by semantic calculations (Smedslund, 1995).

In this way the semantic matrix could pose as a Bayesian prior to research in the social sciences (Gelman and Shalizi, 2013). One can now compute the likely relationships among all variables prior to making empirical data collections (Gefen and Larsen, 2017; Arnulf et al., 2018a; Kjell et al., 2019; Gefen et al., 2020; Rosenbusch et al., 2020; Kjell K. et al., 2021; Nimon, 2021). From a statistical point of view, one should ask questions like who, how, why, and how much people will comply with what is semantically expected.

One study on motivation using semantic analysis found significant differences on individual and group levels in the way that people complied with semantic patterns (Arnulf et al., 2020). Here, different professional group made important group-level deviations from what was semantically expected in a way that correlated highly with the professions' income levels. This calculation involved three data sources with no possible endogeneity: There were the various professions' response patters, combined with the LSA calculated semantics, related to income levels as reported in the national statistics (not from self-report). The numbers strongly suggested that people with higher income levels and education would be directed in their ratings of motivation by a semantic grid that probably matched that of the researchers. People with lower income seemed to twist the meaning of the motivation-related items towards slightly, but significantly different meanings.

In this sense, the semantic grid is not a cast-iron structure. It is probably more like a representational capability with remarkable precision, but not without being malleable in the face of personal and cultural experience. Looking at today's society and challenges in psychology, we are actually faced with challenges to semantic stability at a magnitude that affects political stability. Aside from our psychiatric diagnoses and clinical theories being a source of instability and conflict (Clark et al., 2017), the semantic wars seem to engulf gender, race, political belongingness and perceptions of information trustworthiness (Furnham et al., 2021; Furnham and Horne, 2022).

Given how semantic matrices supply us with experiences of conceptual reality, one should perhaps not wonder that people who are pressed towards conflicts with their own semantic structures will react emotionally, even violently, probably related to what we know about cognitive dissonance (Festinger, 1962; de Vries et al., 2015; Harmon-Jones, 2019). With increasingly powerful computational tools available, we should start describing and outlining the semantic grid to peek at what is behind the horizon.

## The possible neurobiological substrate of the semantic grid

The purpose of the present text is to argue the existence of a semantic representational system that is precise, lean, and not in itself subject to conscious observation. It is possible that this feature of verbal comprehension is founded on a neurobiological correlate. The phenomena we describe are too precise, too independent of culture and too abstract to be the result of local learning processes. Hypothesizing such a mechanism could help to understand its pervasive nature, much like color vision is thought of as a feature of the nervous system and hard to explain to the color blind.

More specifically, we hope to define and identify the mechanism here described as the "semantic grid." Particularly relevant to this pursuit are two arguments explicated by Poeppel et al. (2012) and Krakauer et al. (2017): first, we are trying to delineate this phenomenon so precisely in terms of behavior that a neurobiological substrate could be hypothesized and tested. And second, we describe a function operating on a different level from most of the receptive and productive circuits involved in producing language behavior.

Several lines of studies indicate that the semantic coding of speech content is a specialized function separate from syntactic processing and spanning multiple words (Hagoort and Indefrey, 2014), a process separate from any sensorimotor processing of language. Semantic processing of complex propositions seems associated with a specific area of the medial prefrontal cortex (Hagoort and Indefrey, 2014, p. 354). Concomitantly, world knowledge seems to be treated differently from semantic knowledge in the brain. And finally, the parsing of a literal sentence meaning seems to be a separate step in the process of understanding other people's intentions, indicating that the semantic nature of a proposition is a task on its own, relatively independent of speech acts (Hagoort and Indefrey, 2014, p. 359).

## Summary and conclusion

The purpose of this study has been to review a range of existing empirical publications that use semantic algorithms to predict and model psychological variables and their relationships. We argue that the nature and pervasiveness of semantic predictability should draw attention to how nomological networks can be re-interpreted as semantic networks. Further, we argue that the human capability for processing semantic networks might itself be an important psychological research object, characterized by the following features:

1. It provides a rule-oriented predictability to people's behavior unlike any other behaviors except biological features of the nervous system. This predictability is probably an overlooked, strong law of psychology.
2. It is a measurable structure on an abstract mathematical level that seems to pervade all languages. This feature of the semantic grid provides a measure of similarity in meaning allowing us to translate expressions within and between languages. This points to a biological foundation of the semantic grid that enables culture. It is an open question how much the semantic grid is shaped by culture in return.
3. The computational character of the semantic grid is a constructive feature: It allows survey items and experimental variables to be grouped in accordance with theoretical definitions such that they can be turned into latent constructs. On the other hand, this function constitutes a large matrix that really makes all latent constructs related in some way or other, just like no concepts can exist in isolation from a semantic network. Thus, the "nomological network" argued by Cronbach & Meehl might be determined by (or even be identical to) processing in the semantic grid.
4. One feature of the semantic grid is its automatized character that hides it from conscious experience and hence from psychological investigation. This has led psychological research to adopt a canon for construct validation that locks it in an explanatory room limited upwards to around 42%. Research questions above this threshold will be regarded as same-construct questions. Conversely, for research questions to be argued, they will usually build on semantic networks existing in the semantic grid, driving the *a priori* relationships upwards. The resulting human blindness towards *a priori* relationships is a valid topic for psychological research on its own.
5. The semantic grid does not map truth values. It can only map the mutual meaning of concepts and statements, also for totally fictitious or erroneous ones. It is however sensitive to nonsense.
6. The semantic grid functions as the general matrix within which all definitions and measurement issues take place, forming our epistemic foundations in psychology and creating the "psychological manifest image." We need to recognize and describe it to move past it.
7. The semantic grid is the key standardized communication platform for intersubjective mapping of reality across people. It can be modelled mathematically across subjective experiences as the REC-MTMM model.

## Practical implications

The current state of natural language processing allows researchers to assess how respondents are congruent with the semantic grid. The methodological possibilities are only starting to emerge. For example, it can be used in survey research as follows: At the item level, semantic similarity provides an objective measure that could be used as support for correlating errors in structural equation models. At the scale level, semantic similarity can be used to assess to what degree, if any, empirical nomological networks are based simply on the semantic similarity between item sets (Nimon and Shuck, 2020). In the instrument development stage, semantic similarity can be useful in developing items that are similar to the construct to be measured and divergent from other measures (Rosenbusch et al., 2020; Nimon, 2021). Rather than using eye-ball tests of semantic similarity, researchers can use increasingly available NLP tools to quantify the semantic similarity between two or more item sets or even for automated content validation (Larsen et al., 2023), allowing researchers to quantify discriminant validity.

Studying individual semantic behavior opens the door for future research. In prior research (Arnulf et al., 2018a,b,c,d, 2020; Arnulf and Furnham, 2024), individual semantic acuity or compliance has been shown to be related to personality and cognitive ability. Semantic acuity measures may also be useful in as control variables or to assessing common method variance in lieu of a marker variable, as well as to "unbundle the sample" (Bernardi, 1994, p. 772), identifying subgroups of individuals who yield differential item functioning based on their semantic behavior.

On a more epistemic level, we believe that conceptualizing the semantic grid and its computational properties can help psychology advance to better distinguish between semantically determined and empirically determined discoveries. Semantic computations might be used as a Bayesian priors for separating semantic from empirical relationships.

The rapid development of computerized text analysis and production will probably make text computations as prevalent in the field as factor analysis has been for the recent decades (Arnulf et al., 2021). We believe that psychology can adopt and adapt such tools to make more fruitful distinctions between semantic and empirical questions in the future.

## Limitations

This has been a review of already published studies that use semantic algorithms to predict empirically obtained data patterns. While these studies have found to be predictive of up to around 90% of the observed variation, the claim here is not that all data are semantically determined, nor that the semantic predictions may predict the observed data accurately. The claim is instead that with these possibilities of *a priori* predictions, the nature and meaning of empirical data needs to be considered in light of what is semantically predictable.

Contextual, cultural and statistical factors will always influence the relationships between semantic representations and their observed, empirical counterparts. These influences may be important objects of investigation or disturbing noise depending on the research questions at hand.

Finally, this article has not attempted to make an exhaustive description of how semantic calculations work as it would go beyond the present format. The specific algorithms used and the way the models are designed will affect how the features of the statistics are captured (Arnulf et al., 2018a,b,c,d). However, all reviewed studies contain published descriptions of the technology used. Natural

language processing is rapidly advancing at the time of writing, rendering previously published methods less interesting in the future.

## Author contributions

JA: Conceptualization, Writing – original draft. UO: Writing – original draft, Writing – review & editing. KN: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References


Abdurahman, S., Vu, H., Zou, W., Ungar, L., and Bhatia, S. (2023). A deep learning approach to personality assessment: generalizing across items and expanding the reach of survey-based research. *J. Personal. Soc. Psychol.* doi: 10.1037/pspp0000480

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. *Br. J. Math. Statistic. Psychol.* 49, 347–365. doi: 10.1111/j.2044-8317.1996.tb01093.x

APA (2009). History of the standards. In standards for educational and psychological testing. *Am. Psychol. Assoc.*

Arnulf, J. K., and Furnham, A. (2024). "Never mind the fine print": the interaction of semantics with attitude strength beliefs on corporate cover-ups. *Acta Psychol.* 243:104156. doi: 10.1016/j.actpsy.2024.104156

Arnulf, J. K., and Larsen, K. R. (2018). Cultural insensitivity of Likert-scale surveys in cross-cultural studies of leadership, In Proceeding of AOM meeting, Chicago.

Arnulf, J. K., and Larsen, K. R. (2020). Culture blind leadership research: how semantically determined survey data may fail to detect cultural differences. *Front. Psychol.* 11:176. doi: 10.3389/fpsyg.2020.00176

Arnulf, J. K., Larsen, K., and Dysvik, A. (2018a). Measuring semantic components in training and motivation: a methodological introduction to the semantic theory of survey response. *Hum. Resour. Dev. Q.* 30, 17–38. doi: 10.1002/hrdq.21324

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018b). Respondent robotics: simulating responses to Likert-scale survey items [Vitenskapelig artikkel]. *SAGE Open* 8, 215824401876480–215824401876418. doi: 10.1177/2158244018764803

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018c). Semantic algorithms can detect how media language shapes survey responses in organizational behaviour. *PLoS One* 13, e0207643–e0207626. doi: 10.1371/journal.pone.0207643

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Bong, C. H. (2014). Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS One* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Egeland, T. (2018d). The failing measurement of attitudes: how semantic determinants of individual survey responses come to replace measures of attitude strength. *Behav. Res. Methods* 50, 2345–2365. doi: 10.3758/s13428-017-0999-y

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Nimon, K. F. (2021). Semantic algorithms in the assessment of attitudes and personality. *Front. Psychol.* 12:720559. doi: 10.3389/fpsyg.2021.720559

Arnulf, J. K., Nimon, K., Larsen, K. R., Hovland, C. V., and Arnesen, M. (2020). The priest, the sex worker, and the CEO: measuring motivation by job type. *Front. Psychol.* 11:1321. doi: 10.3389/fpsyg.2020.01321

Avolio, B. J., Bass, B. M., and Jung, D. I. (1995). *Multifactor leadership questionnaire technical report*. Redwood City, CA: Mind Garden.

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: methodological and philosophical foundations. *MIS Q.* 35, 261–292. doi: 10.2307/23044044

Bagozzi, R. P., and Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organ. Res. Methods* 1, 45–87. doi: 10.1177/109442819800100104

Behr, D., Braun, M., and Dorer, B. (2016). Measurement instruments in international surveys. GESIS Survey Guidelines. Available at: https://www.gesis.org/fileadmin/upload/SDMwiki/BehrBraunDorer_Measurement_Instruments_in_Cross-National_Surveys.pdf

Bernardi, R. A. (1994). Validating research results when Cronbach's alpha is below. 70: a methodological procedure. *Educ. Psychol. Meas.* 54, 766–775. doi: 10.1177/0013164494054003023

Blanchette, P. A. (2012). *Frege's conception of logic*. Oxford, UK: Oxford University Press.

Borge-Holthoefer, J., and Arenas, A. (2010). Semantic networks: structure and dynamics. *Entropy* 12, 1264–1302. doi: 10.3390/e12051264

Boring, E. G. (1945). The use of operational definitions in science. *Psychol. Rev.* 52, 243–245. doi: 10.1037/h0054934

Borsboom, D. (2008). Latent variable theory. *Measurement* 6, 25–53. doi: 10.1080/15366360802035497

Bridgman, P. W. (1927). *The logic of modern physics*. New York, NY: Macmillan.

Brunila, M., and LaViolette, J. (2022). What company do words keep? Revisiting the distributional semantics of JR Firth & Zellig Harris. *arXiv preprint arXiv*:2205.07750. doi: 10.48550/arXiv.2205.07750

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Clark, L. A., Cuthbert, B., Lewis-Fernandez, R., Narrow, W. E., and Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's research domain criteria (RDoC). *Psychol. Sci. Public Interest* 18, 72–145. doi: 10.1177/1529100617727266

Cohen, J. R., Swerdlik, M. E., and Sturman, E. D. (2013). *Psychological testing and assessment: an introduction to tests and measurement*. *8th* Edn. New York, NY: McGraw-Hill.

Colquitt, J. A., Sabey, T. B., Rodell, J. B., and Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *J. Appl. Psychol.* 104, 1243–1265. doi: 10.1037/apl0000406

Coombs, C. H. (1964). *A theory of data*. New York, NY: Wiley.

Coombs, C. H., and Kao, R. C. (1960). On a connection between factor-analysis and multidimensional unfolding. *Psychometrika* 25, 219–231. doi: 10.1007/Bf02289726

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

de Vries, J., Byrne, M., and Kehoe, E. (2015). Cognitive dissonance induction in everyday life: an fMRI study [article]. *Soc. Neurosci.* 10, 268–281. doi: 10.1080/17470919.2014.990990

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407. doi: 10.1002/(Sici)1097-4571(199009)41:6<391::Aid-Asi1>3.0.Co;2-9

Dennett, D. (2012). A perfect and beautiful Machine': what Darwin's theory of evolution reveals about artificial intelligence. The Atlantic. Available at: https://www.theatlantic.com/technology/archive/2012/06/-a-perfect-and-beautiful-machine-what-darwins-theory-of-evolution-reveals-about-artificial-intelligence/258829/

Dennett, D. C. (2013). Bestiary of the manifest image. *Sci. Metaph.* 96, 96–107. doi: 10.1093/acprof:oso/9780199696499.003.0005

Dennett, D. (2018). *From bacteria to Bach and back: the evolution of minds*. London: Penguin Books.

Dennis, S., Landauer, T. K., Kintsch, W., and Quesada, J. (2013). *Introduction to latent semantic analysis*. Boulder, CO: University of Colorado.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805 2018.

Drasgow, F., Chernyshenko, O. S., and Stark, S. (2015). 75 years after Likert: Thurstone was right! *Ind. Organ. Psychol.* 3, 465–476. doi: 10.1111/j.1754-9434.2010.01273.x

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In Conference on human factors in computing systems: Proceedings of the SIGCHI conference on human factors in computing systems, Washington, D.C.

Festinger, L. (1962). Cognitive dissonance. *Sci. Am.* 207, 93–106. doi: 10.1038/scientificamerican1062-93

Firth, J. R. (1957). "A synopsis of linguistic theory, 1930-1955", in *Studies in linguistic analysis, Special Volume of the Philological Society*, Blackwell, 1–31.

Frege, G. (1918). "Der Gedanke. Eine logische Untersuchung" in *Beiträge zur Philosophie des deutschen Idealismus I*. Erfurt: Deutsche Philosophische Gesellschaft, 58–77.

Freiberg, B., and Matz, S. C. (2023). Founder personality and entrepreneurial outcomes: a large-scale field study of technology startups. *Proc. Natl. Acad. Sci.* 120:e2215829120. doi: 10.1073/pnas.2215829120

Furnham, A., Arnulf, J. K., and Robinson, C. (2021). Unobtrusive measures of prejudice: estimating percentages of public beliefs and behaviours. *PLoS One* 16:e0260042. doi: 10.1371/journal.pone.0260042

Furnham, A., and Horne, G. (2022). Cover ups and conspiracy theories: demographics, work disenchantment, equity sensitivity, and beliefs in cover-ups. *J. Work Organ. Psychol.* 38, 19–25. doi: 10.5093/jwop2022a2

Fyffe, S., Lee, P., and Kaplan, S. (2023). "Transforming" personality scale development: illustrating the potential of state-of-the-art natural language processing. *Organ. Res. Methods*:10944281231155771. doi: 10.1177/10944281231155771

Gefen, D., Fresneda, J. E., and Larsen, K. R. (2020). Trust and distrust as artifacts of language: a latent semantic approach to studying their linguistic correlates. *Front. Psychol.* 11:561. doi: 10.3389/fpsyg.2020.00561

Gefen, D., and Larsen, K. (2017). Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. *J. Assoc. Inf. Syst.* 18, 727–757. doi: 10.17705/1jais.00469

Gelman, A., and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66, 8–38. doi: 10.1111/j.2044-8317.2011.02037.x

Haeffel, G. J. (2022). Psychology needs to get tired of winning. *R. Soc. Open Sci.* 9:220099. doi: 10.1098/rsos.220099

Hagoort, P., and Indefrey, P. (2014). The neurobiology of language beyond single words. *Annu. Rev. Neurosci.* 37, 347–362. doi: 10.1146/annurev-neuro-071013-013847

Harari, Y. N. (2015). *Sapiens: a brief history of humankind*. New York, NY: Harper Collins.

Harmon-Jones, E. E. (2019). *Cognitive dissonance: Reexamining a pivotal theory in psychology*. Washington, DC: American Psychological Association.

Hergenhahn, B. R. (2009). *An introduction to the history of psychology*. *6th* Edn. Belmont, CA: Wadsworth Cengage Learning.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organ. Res. Methods* 1, 104–121. doi: 10.1177/109442819800100106

Hinkin, T. R., and Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organ. Res. Methods* 2, 175–186. doi: 10.1177/109442819922004

Jöreskog, K. G. (1993). "Testing structural equation models" in *Testing structural equation models*. eds. K. A. Bollen and J. S. Long (Newbury Park, CA: Sage), 294–316.

Kahneman, D., and Tversky, A. (1982). The psychology of preferences. *Sci. Am.* 246, 160–173. doi: 10.1038/scientificamerican0182-160

Kjell, O., Daukantaitė, D., and Sikström, S. (2021). Computational language assessments of harmony in life — not satisfaction with life or rating scales — correlate with cooperative behaviors. *Front. Psychol.* 12:601679. doi: 10.3389/fpsyg.2021.601679

Kjell, K., Johnsson, P., and Sikström, S. (2021). Freely generated word responses analyzed with artificial intelligence predict self-reported symptoms of depression, anxiety, and worry. *Front. Psychol.* 12:602581. doi: 10.3389/fpsyg.2021.602581

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikstrom, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., and Den Hartog, D. N. (2018). Text mining in Organizational Research. *Organ. Res. Methods* 21, 733–765. doi: 10.1177/1094428117722619

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., Mac Iver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Kyngdon, A. (2006). An introduction to the theory of unidimensional unfolding. *J. Appl. Meas.* 7, 260–277.

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: what is it? How was it contracted? Is there a cure? *New Ideas Psychol.* 31, 65–71. doi: 10.1016/j.newideapsych.2011.02.009

Landauer, T. K. (2007). "LSA as a theory of meaning" in *Handbook of latent semantic analysis*. eds. T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsh (Mahwah, NJ: Larence Erlbaum Associates, Publishers), 3–34.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037//0033-295x.104.2.211

Landauer, T. K., and Littman, M. L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In Proceedings of the sixth annual conference of the UW Centre for the new Oxford English dictionary and text Research, Waterloo, ON, Canada.

Lange, R., Greyson, B., and Houran, J. (2015). Using computational linguistics to understand near-death experiences: concurrent validity for the near death experience scale. *Psychol. Conscious. Theory Res. Pract.* 2, 79–89. doi: 10.1037/cns0000040

LaPiere, R. T. (1934). Attitudes vs. actions. *Soc. Forces* 13, 230–237. doi: 10.2307/2570339

Larsen, K. R., and Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and Meta-analyses. *MIS Q.* 40, 529–551. doi: 10.25300/MISQ/2016/40.3.01

Larsen, K. R., Sharma, R., Quieroz, M., Arnulf, J. K., and Pillet, J.-C. (2023) Use of natural language processing techniques in the construct and instrument development process [Konferanse]. In proceeding of the Thirty-first European conference on information systems (ECIS 2023), Kristiansand: ECIS, European Conference on Information Systems.

Larsen, K. R., Voronovich, Z. A., Cook, P. F., and Pedro, L. W. (2013).Addicted to constructs: science in reverse? *Addiction (Abingdon, England)* 108, 1532–1533.

Lovasz, N., and Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. *New Ideas Psychol.* 31, 22–31. doi: 10.1016/j.newideapsych.2011.02.005

Mac Kenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. (2011). Construct measurement and validation procedures in Mis and behavioral research: integrating new and existing techniques [article]. *MIS Q.* 35, 293–334. doi: 10.2307/23044045

Mari, L., Maul, A., Irribarra, D. T., and Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement* 100, 115–121. doi: 10.1016/j.measurement.2016.12.050

Martinsen, Ø. L., Arnulf, J. K., Larsen, K. R., Ohlsson, U. H., and Satorra, A. (2017). Semantic influence on the measurement of leadership: a multitrait-multisource perspective. Paper Presented at the Academy of Management Meeting, Atlanta

McGrane, J. A., and Maul Gevirtz, A. (2019). The human sciences, models and metrological mythology. *Measurement* 152:107346. doi: 10.1016/j.measurement.2019.107346

Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X10000968

Mercier, H., and Sperber, D. (2019). *The enigma of reason*. Boston, MA: Harvard University Press.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *J. Math. Psychol.* 38, 244–273. doi: 10.1006/jmps.1994.1016

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas Psychol.* 31, 13–21. doi: 10.1016/j.newideapsych.2011.02.004

Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., and Zweig, G. (2013). Word 2vec. Available at: https://code.google.com/p/word2vec

Nimon, K. F. (2021). MOWDOC: a dataset of documents from taking the measure of work for building a latent semantic analysis space [data report]. *Front. Psychol.* 11:523494. doi: 10.3389/fpsyg.2020.523494

Nimon, K., and Shuck, B. (2020). Work engagement and burnout: testing the theoretical continuums of identification and energy. *Hum. Resour. Dev. Q.* 31, 301–318. doi: 10.1002/hrdq.21379

Nimon, K., Shuck, B., and Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: a function of semantic equivalence? [article]. *J. Happiness Stud.* 17, 1149–1171. doi: 10.1007/s10902-015-9636-6

Northouse, P. G. (2021). *Leadership: theory and practice*. Thousand Oaks, CA: Sage Publications.

Nunally, J. C., and Bernstein, I. H. (2007). *Psychometric theory*. *3rd* Edn. New York, NY: McGraw-Hill.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociol. Methods Res.* 27, 226–284. doi: 10.1177/0049124198027002004

Pearl, J. (2009). *Causality*. New York, NY: Cambridge University Press.

Pearl, J., and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. New York, NY: Basic Books.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58, 240–242.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* 60, 489–498.

Pierce, J. R. (1980). *Introduction to information theory: symbols, signals and noise. 2nd* Edn. New York, NY: Dover Publications, Inc.

Pillet, J.-C., Larsen, K. R., Queiroz, M., Arnulf, J. K., and Sharma, R. (2022). L'Effet Perturbateur du Nom des Construits sur les Conclusions aux Tests de Validité de Contenu. In Résultats Préliminaires et Interprétation 27e Conférence de l'Association Information et Management (AIM)

Poeppel, D., Emmorey, K., Hickok, G., and Pylkkänen, L. (2012). Towards a new neurobiology of language. *J. Neurosci.* 32, 14125–14131. doi: 10.1523/jneurosci.3244-12.2012

Popper, K. (1959). *The logic of scientific discovery*, vol. *12*. New York, NY: Basic Books, 53–54.

Popper, K., and Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature* 302, 687–688. doi: 10.1038/302687a0

Proix, T., Saa, J. D., Christen, A., Martin, S., Pasley, B. N., Knight, R. T., et al. (2021). Imagined speech can be decoded from low-and cross-frequency features in perceptual space. *bioRxiv* 2021:428315. doi: 10.1101/2021.01.26.428315

Rosenbusch, H., Wanders, F., and Pit, I. L. (2020). The semantic scale network: an online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychol. Methods* 25, 380–392. doi: 10.1037/met0000244

Russell, B. (1918/2007). "The relation of sense-data to physics" in *Mysticism and logic*. ed. B. Russell (Nottingham, UK: Bertrand Russell Peace Foundation), 139–170.

Russell, B. (1919). *Introduction to mathematical philosophy*. London, UK: George Allen & Unwin.

Russell, B. (1922). "An introduction to the tractatus logico-philosophicus" in *Tractatus logico-philosophicus*. ed. L. Wittgenstein (London, UK: Kegan Paul)

Ryle, G. (1937). Categories. *Proc. Aristot. Soc.* 189–206.

Satorra, A., Olsson, U. H., Arnulf, J. K., and Martinsen, Ø. L. (2023). The REC-MTMM (a conference publication.) The REC-MTMM Johan Arndt conference, Bergen, Norway.

Semin, G. (1989). The contribution of linguistic factors to attribute inference and semantic similarity judgements. *Eur. J. Soc. Psychol.* 19, 85–100. doi: 10.1002/ejsp.2420190202

Shannon, C., and Weaver, W. (1949). *The mathematical theory of communication*. Urbana, Illinois: University of Illinois Press.

Shuck, B., Nimon, K., and Zigarmi, D. (2017). Untangling the predictive Nomological validity of employee engagement: partitioning variance in employee engagement using job attitude measures. *Group Org. Manag.* 42, 79–112. doi: 10.1177/1059601116642364

Slaney, K. L. (2017a). (Ed.) "Some conceptual housecleaning" in *Validating psychological constructs: Historical, philosophical, and practical dimensions* (London, UK: Palgrave Macmillan UK), 201–234.

Slaney, K. L. (2017b). (Ed.) *Validating psychological constructs: historical, philosophical, and practical dimensions.* London, UK: Palgrave Mac Millan.

Slaney, K. L., and Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas Psychol.* 31, 4–12. doi: 10.1016/j.newideapsych.2011.02.003

Smedslund, J. (1994). Nonempirical and empirical components in the hypotheses of 5 social-psychological experiments [article]. *Scand. J. Psychol.* 35, 1–15. doi: 10.1111/j.1467-9450.1994.tb00928.x

Smedslund, J. (1995). Psychologic: commonsense and the pseudoempirical. In J. Smith, R. Harre and LangenhoveL. Van (Eds.), *Rethinking psychology* 196–206. London, UK: Sage.

Smedslund, J. (2012). Psycho-logic: some thoughts and after-thoughts. *Scand. J. Psychol.* 53, 295–302. doi: 10.1111/j.1467-9450.2012.00951.x

Smedslund, G., Arnulf, J. K., and Smedslund, J. (2022). Is psychological science progressing? Explained variance in Psycinfo articles during the period 1956 to 2022. *Front. Psychol.* 13:1089089. doi: 10.3389/fpsyg.2022.1089089

Thorndike, E. (1904). *An introduction to the theory of mental and social measurements*. Hoboken, NJ: Teachers College.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Uher, J. (2021a). Functions of units, scales and quantitative data: fundamental differences in numerical traceability between sciences. *Qual. Quant.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2021b). Quantitative psychology under scrutiny: measurement requires not result-dependent but traceable data generation. *Personal. Individ. Differ.* 170:110205. doi: 10.1016/j.paid.2020.110205

van Knippenberg, D., and Sitkin, S. B. (2013). A critical assessment of charismatic—transformational leadership research: Back to the drawing board? *Acad. Manag. Ann.* 7, 1–60. doi: 10.1080/19416520.2013.759433

Vessonen, E. (2019). Operationalism and realism in psychometrics. *Philos Compass* 14:e12624. doi: 10.1111/phc3.12624

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London, UK: Kegan Paul, Trench, Trubner & CO., Ltd.

Wittgenstein, L. (1953). *Philosophische Untersuchungen*. Oxford, UK: Basil Blackwell.

frontiers | Frontiers in Psychology

TYPE Perspective
PUBLISHED 27 March 2024
DOI 10.3389/fpsyg.2024.1389640



OPEN ACCESS

EDITED BY
Barbara Hanfstingl,
University of Klagenfurt, Austria

REVIEWED BY
Abdolvahab Khademi,
University of Massachusetts Amherst,
United States

*CORRESPONDENCE
David Carré
✉ david.carre@uoh.cl

RECEIVED 21 February 2024
ACCEPTED 14 March 2024
PUBLISHED 27 March 2024

CITATION
Paredes J and Carré D (2024) Looking for a broader mindset in psychometrics: the case for more participatory measurement practices.
*Front. Psychol.* 15:1389640.
doi: 10.3389/fpsyg.2024.1389640



# Looking for a broader mindset in psychometrics: the case for more participatory measurement practices

Javiera Paredes[1] and David Carré[2]*

[1]Laboratorio de Lenguaje, Interacción y Fenomenología, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile, [2]Instituto de Ciencias de la Salud, Universidad de O'Higgins, Rancagua, Chile

Psychometrics and the consequences of its use as *the* method of quantitative empirical psychology has been continuously criticized by both psychologists and psychometrists. However, the scope of the possible solutions to these issues has been mostly focused on the establishment of methodological-statistical best practices for researchers, without any regard to the pitfalls of previous stages of measurement as well as theory development of the targeted phenomenon. Conversely, other researchers advance the idea that, since psychometrics is riddled with many issues, the best way forward is a complete rework of the discipline even if it leaves psychologists and other practitioners without any way to measure quantitatively for a long period of time. Given these tensions, we therefore advocate for an alternative path to consider while we work on making substantive change in measurement. We propose a set of research practices focusing on the inclusion and active participation of groups involved in measurement activities, such as psychometrists, researchers but most importantly practitioners and potential participants. Involving a wider community while measuring in psychology could tackle some key issues that would take us closer to a more authentic approach to our phenomenon of interest.

KEYWORDS

**psychometrics, measurement, psychology, participation, communities**

## Introduction

By looking at the current landscape of psychology, there are many reasons to argue that psychometrics is one of the most successful subfields of the discipline (Borsboom and Wijsen, 2017; Craig, 2017). It is cited and used by almost every empirical work published in recent decades (Jones and Thissen, 2006). Even more so, its measurement standards have become basic requisites asked by most scientific journals to even consider a manuscript for review (Eich, 2014; Trafimow and Marks, 2015). Accordingly, it has become a–if not *the*–core course of almost every undergraduate and graduate program in any field related to psychological science (Friedrich et al., 2000; TARG Meta-Research Group, 2022). In brief, contemporary psychological research seems to involve putting psychometrics into practice.

Considering its success and widespread influence, it is nothing short of paradoxical that psychometrics has been the target of the harshest critiques within and beyond the discipline during recent decades. The range of these critiques has gone from questioning whether the last 50 years of psychometric research has any value at all (Salzberger, 2013)

to arguing that psychometrics does not actually do measurement–at least in the metrological sense of the term (Uher, 2021a,b). Thus, psychometrics has been criticized to its core, ultimately calling for its refoundation.

Even if we look past these fundamental critiques, we find that researchers within the psychometrics community have also raised a number of issues; which they have tried to address with varying degrees of success. Among these it is possible to find the replicability crisis (Stevens, 2017; Anvari and Lakens, 2018), all sorts of data dredging practices, commonly known as *p-hacking* (Szucs, 2016; Stefan and Schönbrodt, 2023), or the lack of pre-registering protocols (van 't Veer and Giner-Sorolla, 2016; Spitzer and Mueller, 2023). While the latter group of critiques has called for necessary improvements of standards and practices within psychometrics and psychology, they have not really addressed the breadth and depth of the criticisms made by other scholars (e.g., Salzberger, 2013; Uher, 2021b). Neither have they tried to: during the last decade most of the psychometric community have been devoted to developing procedures and practices aimed to prevent the misuse of psychometrics by researchers. Thus, the effort to solve the aforementioned issues has focused on turning detailed data-handling protocols and replication studies into common practices within psychological research. But they do not question—with exceptions (e.g., Bauer, 2024)–whether psychometrics actually measures what it aims to measure or even if it measures something at all. This second group of critiques thus follows a line of renovating psychometrics rather than rebuilding it. This, in turn, makes the dialog between both camps unlikely: as one side aims to change (almost) everything from the ground up while the other looks to correct and prevent malpractices.

In this scenario, the present work neither aims to deepen the re-foundational critiques that have been posed on psychometrics, nor proposes adjustments to current measurement practices hoping to solve all the ailments of the discipline. Instead, we aim to build upon already identified issues to propose alternative *research* practices for psychometrics that broaden the mindset of this sub discipline. We argue that these practices could contribute in closing the gap between existing critiques and the current measurement standards in a feasible way.

We do so for two reasons. First, despite the recognition of its many shortcomings and the conceptual critiques against its tenets and practices, psychometrics keeps—and probably will keep—being utilized by practitioners and researchers alike due to its standing and usefulness. Thus, the prospect of rebuilding the discipline, starting something new based upon completely different tenets, seems simply unfeasible. Second, because we do acknowledge that changes in psychometric practices have to go beyond pre-registering, statistical and open data practices. In order to make changes to psychometrics substantial, they have to alter the direction in which current research and measurement practices are pointed. This is why we consider that more transparency, expressed through different procedures (e.g., Hardwicke and Vazire, 2023), is not enough by itself to make psychometrics–and its impact over psychological research and practice at large–overcome its fundamental challenges.

For these reasons, what we deem essential is a change in the mindset of psychometrics toward a broader one. A change that does not aim to make psychometrics renounce to technical and mathematical standards (which would be an oxymoron), but not to make these standards its only interest and ultimate goal. We are not alone in proposing a change of this kind. In a recent editorial, the outgoing editor-in-chief of *Psychological Science*–one of the journals with highest impact factor in psychology–calls for a similar change: to stop focusing all the attention on methodological, procedural issues and start thinking about how psychological research actually speaks about the phenomena of interest, which she aptly terms as authenticity (Bauer, 2024). We share with Bauer (2024) that doing *more* is not enough, it has to be done *differently*.

In the following we argue in favor of a set of practices that could–and should–be done differently: participatory processes within measurement practices. More specifically, we focus on the role that promoting participation could have on achieving a better understanding of the measurement processes involved in the most common psychometric instruments–namely, questionnaires (see Tourangeau et al., 2000). As it has been proposed (Uher, 2021b), the person being the instrument of measurement is one of the essential shortcomings of psychometrics. We consider that, for a discipline devoted to human-driven measurement, this is rather one of the essential challenges of psychometrics.

## Humans as data generation instruments

One of the fundamental issues identified by critics of psychometrics focuses on the human-based nature of measurement in psychology (Uher, 2021a,b). Since the use of surveys in psychology is extensive, the participant–as defined by metrology–is regarded as the source of the quantitative data. It is the person who reads, understands and interprets the instrument the one to give an answer related to the construct that the survey ultimately refers to (Uher, 2021b). Different to this response process is the structure of the scale itself. Scales may or may not follow different psychometric standards, which is determined by the statistical analysis of the numerical responses that were provided by human action.

The metrological perspective, however, is in clear opposition to what psychology typically considers as the source of data generated by quantitative instruments. In the common use of psychometrics by psychologists and practitioners, the measurement instrument is determined by the number of questionnaire items defined as latent representatives of the studied phenomena. The participants who respond to the survey are not usually considered primary players in the response process beyond providing data for validation processes during measurement development (Hughes, 2018; Levac et al., 2019; Reynolds et al., 2021). Therefore, after validation, instruments seem to gain a life of its own that transcends the way in which respondents interact with them.

This naïve approach to quantitative measurement involving instruments such as surveys in psychology implies a double source of possible error. Participants, according to this view, produce an answer to the latent construct that the survey asks for. But the former neglects that the construction of the items already has an identified source of error, which stems from the distance between the particular construct proposed by instrument-developers and the theoretical definition of the psychological concept that encompasses all its possible modes of presentation (Uher, 2018). This first source of error, namely the distance between the construct and its theoretical definition, has been long identified by psychologists through the empirical testing of their

measurement models. Researchers have long discussed the inability of quantitative psychological models to achieve complete fidelity to the phenomena studied through the developed measurement instruments (Oberauer and Lewandowsky, 2019; Eronen and Bringmann, 2021).

The second possible source of error emerges every time that a particular participant answers each item of the survey. How do we know that the cognitive and interpretative process is the same in every person that approaches the instrument? This well-known issue is commonly addressed in the process of developing measures through tools like the cognitive interview. This interview aims to figure out the response processes to make sure that each item is understood as the researchers intended it to (Tourangeau et al., 2000). This approach, however, does not solve the fact that each singular process of response could bring very different outcomes by the only act of interpretation of each participant. For example, how does a headache affect the process of understanding what happiness is? Contextual elements, beyond the cumulative of cognitive representational contents assigned to each definition of an item during validation, could be an inextricable source of error related to the human-based nature of measurement in psychology.

To summarize, the measurement process in psychology relies on two different user-dependent activities: one that involves the appropriate understanding of the scale functioning by researchers and practitioners; and the agreement of each person on the definition of the phenomena presented as items in the questionnaires. It is in this regard that person-centered interactions and instruments are considered by metrologists as one the roots of measurement errors in psychological assessment. Numerical traceability is one of the critical aims in quantitative measurement to ensure a successful data generation process. Successful, in this context, implies the existence of a clear link between the numerical attributes assigned to psychological phenomena and certain pre-established standards. For a link that directly relates the numerical attribute with the psychological phenomena is the only way to make results obtained from questionnaires to be non-dependent on the users of the instrument (Uher, 2021b). Therefore, when we consider the human-based nature of measurement described above, numerical traceability in psychology is not achievable.

The recommendation of experts when confronted with the issue of the lack of numerical traceability in psychology has been to search for practices to ensure the establishment of clear and distinct *intersubjective* meanings of the numerical results of each item (e.g., Hughes, 2018; Reynolds et al., 2021). A successful example of these practices is identified in the development of cognitive abilities instruments such as the Wechsler Adult Intelligence Scale (WAIS) (Benson et al., 2010; Weiss et al., 2010). Since the process of cognitive evaluation has an additional human-based source of error (i.e., the test applicator) and the stakes involved in this kind of assessment process are high, the need of establishing clear meanings regarding numerical results is just as key as the conceptual nature of the constructs evaluated. The results of these practices are certainly satisfactory, as the meaning of numerical results of the WAIS are fairly standard and unambiguous within the cognitive assessment community.

Here it is important to note that we see no contradiction between, on the one hand, improving measurement practices in order to provide an account of the phenomena that is closer to the theoretical grounds proposed and, on the other, advancing toward more precise theoretical structures that allow numerical traceability in psychology. Therefore, we follow the experts' recommendation and further argue that there is much to be gained in attempting to make conventional, intersubjective agreements about numerical results more common across the discipline; for examples like the one described above are the exception rather than the norm. To do so, as we develop in the following, it is essential to involve actors beyond psychometrics to make such intersubjective agreements actually agreements and not yet another technical recommendation.

## Participatory processes as a cornerstone of psychological measurement

As we argued at the beginning of this work, we consider that psychometrics is in dire need of broadening its mindset. By this we mean that rather than trying to do more—or less—of what is currently done, different things should be done instead. Thinking along these lines, we are in favor of promoting community participatory processes as a pivotal element of measurement practices in psychology. By community participatory processes we are standing for the inclusion of researchers, practitioners and users of psychological instruments.

As noted above, the inclusion of best practices in psychological research and publication has been the cornerstone of the attempts to solve the issues regarding measurement in psychology (e.g., Flake and Fried, 2020; Aguinis et al., 2021). Naturally, the community involved in these changes has mostly included psychometrists and researchers in psychology. We believe, however, that the efforts toward improving measurement instruments should also involve the voices of more practitioners and everyday users of these instruments, even–or especially–if they are not trained in psychological science.

Practitioners and users of the instruments developed by psychometrists and researchers are essential stakeholders that possess insights into some pressing issues in this discussion, like numeric traceability. Achieving agreement about the intersubjective meaning of scale items is one example, as described above. An accurate analysis of these problems only can be conducted when the developers of the instruments can account for the understanding of all the people involved in these practices. Users and practitioners, therefore, should not only be eventually included in the process in the final stages of development (i.e., validation) but also in previous steps, thus assisting the construction of measurements that are sensible to the phenomena of interest.

Respondents, on the other hand, are a source of crucial information regarding the actual interpretation and response processes in surveys. While we may rely on the expertise of psychologists, psychometrists and, sometimes, the teams that apply these instruments, it is not enough to capture the real meaning given by people to each item. And the main issue still remains intact if we consider that we as psychologists still rely heavily on samples that do not necessarily represent the people who answer our surveys. WEIRD (western, educated, industrialized, rich, and democratic) or Mechanical Turk samples have been the focus of past and current academic discussion regarding their suitability as a source of data in psychological research (Keith and Harms, 2016; Webb and Tangney, 2022).

A fair counterpoint to more participatory practices is the issue of viability. The inclusion of every single prospective practitioner or user, and including each meaning considered to the item construction and instrument it is simply not achievable, especially when means are scarce and time is limited. But that would be taking the argument to an unreasonable extreme. What we are proposing here is making efforts for a wider and more nuanced understanding of how different people, communities and cultures approach and answer the scales that are developed. Participation is anything but binary, thus we are calling for advancing toward more inclusion of different actors and not for a strict process of co-creation.

Once again, the way in which cognitive assessment has included participatory practices offers valuable insights. Even without modifying the instruments used, this area has shown how to improve existent measurement practices in psychology. Due to the practical impact that such an assessment has, it commonly involves lengthy validation efforts that ensure that the data generation instruments–namely, people–are participating and responding in such a way that can be compared to other persons in other areas of the world. But the stakes of psychological measurement certainly go beyond cognitive assessment. Determining levels of prejudice among members of a community; assessing whether a person meets a specific personality profile; establishing the impact of an intervention in the improvement of memory. These examples, as many others do, remind us of the stakes involved in developing psychometric instruments. They should also push us to make every possible effort to improve measurement practices–even if it involves costlier and slower development processes that include participation.

## The siren's call for quick data collection

In this perspective work we have argued in favor of expanding the current mindset of psychometrics in order to look beyond technical and statistical concerns. We do so to advance potential solutions to the pressing challenges of the subdiscipline without waiting for its refoundation or hoping for minor renovations. Although a complex endeavor, we cannot ignore precisely what makes psychological measurement prone to error, the human-based nature of the data-generation instrument.

Instead of trying to look past this human nature through sophisticated means, we have proposed ways to understand this nature better through participatory practices. Therefore, the psychometric and psychological communities of researchers should not disregard the attitudes, meanings and knowledge of other groups involved in measurement–that is if they want to develop instruments that account for the complex psychological phenomena they measure.

These ideas, moreover, could also be applied to measurement in other disciplines in which participation has not been a priority. In educational assessment, a number of works have emphasized participation mostly through self- and peer-assessment practices (e.g., Li et al., 2016) and teacher's practices for communicating their assessment expectations (e.g., Stefani, 1998). In standardized testing, the general absence of participatory practices should not come as a surprise considering that the *Standards for Educational And Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) mentions 'participatory' only once in its 130 pages. Therefore, participation has been reduced to processes that do not actually involve students on what or how their learning is assessed (Aarskog, 2021). In health sciences, on the other hand, there is devoted effort to enhance user's participation in multiple dimensions of healthcare (Angel and Frederiksen, 2015); except in the development of instruments used to assess health outputs. In sum, we envision a significant space for including the practices processes we propose, although the specific way in which different fields could bring these ideas into everyday practice, however, remains an open discussion that we hope to trigger with this work.

We have no doubts that our position does not sit well with many researchers in psychometrics who honestly hope to address every single issue through technical means. To them, we can only repeat the blunt conclusion of Patricia Bauer's recent editorial piece: "(...) we must resist the siren's call for quick data collection, with instruments that barely scratch the surface of a complex psychological construct, and that offer sweeping conclusions seemingly without limits on their generalizability." (2024, p.3) One of the ways in which we can resist that call is bringing more voices into the work of psychometrics and make them participate in the development of psychological measurement.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

JP: Conceptualization, Writing – original draft, Writing – review & editing. DC: Conceptualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. DC was supported by ANID-FONDECYT Postdoctorado (Grant no. 3200593).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

# References


Aarskog, E. (2021). 'No assessment, no learning': exploring student participation in assessment in Norwegian physical education (PE). *Sport Educ. Soc.* 26, 875–888. doi: 10.1080/13573322.2020.1791064

Aguinis, H., Hill, N. S., and Bailey, J. R. (2021). Best practices in data collection and preparation: recommendations for reviewers, editors, and authors. *Organ. Res. Methods* 24, 678–693. doi: 10.1177/1094428119836485

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angel, S., and Frederiksen, K. N. (2015). Challenges in achieving patient participation: a review of how patient participation is addressed in empirical studies. *Int. J. Nurs. Stud.* 52, 1525–1538. doi: 10.1016/j.ijnurstu.2015.04.008

Anvari, F., and Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Compr. Results Soc. Psychol.* 3, 266–286. doi: 10.1080/23743603.2019.1684822

Bauer, P. J. (2024). Attention to authenticity: an essential analogue to focus on rigor and replicability. *Psychol. Sci.* 35, 3–6. doi: 10.1177/09567976231220895

Benson, N., Hulac, D. M., and Kranzler, J. H. (2010). Independent examination of the Wechsler adult intelligence scale—fourth edition (WAIS-IV): what does the WAIS-IV measure? *Psychol. Assess.* 22, 121–130. doi: 10.1037/a0017767

Borsboom, D., and Wijsen, L. D. (2017). Psychology's atomic bomb. *Assess. Educ.* 24, 440–446. doi: 10.1080/0969594X.2017.1333084

Craig, K. (2017). "The history of psychometrics" in *Psychometric testing*. ed. K. Craig (New York: John Wiley & Sons, Ltd.), 1–14.

Eich, E. (2014). Business not as usual. *Psychol. Sci.* 25, 3–6. doi: 10.1177/0956797613512465

Eronen, M. I., and Bringmann, L. F. (2021). The theory crisis in psychology: how to move forward. *Perspect. Psychol. Sci.* 16, 779–788. doi: 10.1177/1745691620970586

Flake, J. K., and Fried, E. I. (2020). Measurement Schmeasurement: questionable measurement practices and how to avoid them. *Adv. Methods Pract. Psychol. Sci.* 3, 456–465. doi: 10.1177/2515245920952393

Friedrich, J., Buday, E., and Kerr, D. (2000). Statistical training in psychology: a National Survey and commentary on undergraduate programs. *Teach. Psychol.* 27, 248–257. doi: 10.1207/S15328023TOP2704_02

Hardwicke, T. E., and Vazire, S. (2023). Transparency is now the default at psychological science. *Psychol. Sci.* 2023:1573. doi: 10.1177/09567976231221573

Hughes, D. J. (2018). "Psychometric validity" in *The Wiley handbook of psychometric testing* (New York: John Wiley & Sons, Ltd), 751–779.

Jones, L. V., and Thissen, D. (2006). "1 a history and overview of psychometrics" in *Handbook of statistics*. eds. C. R. Rao and S. Sinharay, vol. *26* (Amsterdam, Netherlands: Elsevier), 1–27.

Keith, M. G., and Harms, P. D. (2016). Is mechanical Turk the answer to our sampling woes? *Ind. Organ. Psychol.* 9, 162–167. doi: 10.1017/iop.2015.130

Levac, L., Ronis, S., Cowper-Smith, Y., and Vaccarino, O. (2019). A scoping review: the utility of participatory research approaches in psychology. *J. Community Psychol.* 47, 1865–1892. doi: 10.1002/jcop.22231

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., et al. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assess. Eval. High. Educ.* 41, 245–264. doi: 10.1080/02602938.2014.999746

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* 26, 1596–1618. doi: 10.3758/s13423-019-01645-2

Reynolds, C. R., Altmann, R. A., and Allen, D. N. (2021). "Validity" in *Mastering modern psychological testing: Theory and methods*. eds. C. R. Reynolds, R. A. Altmann and D. N. Allen (Berlin: Springer International Publishing), 185–222.

Salzberger, T. (2013). Attempting measurement of psychological attributes. *Front. Psychol.* 4:75. doi: 10.3389/fpsyg.2013.00075

Spitzer, L., and Mueller, S. (2023). Registered report: survey on attitudes and experiences regarding preregistration in psychological research. *PLoS One* 18:e0281086. doi: 10.1371/journal.pone.0281086

Stefan, A. M., and Schönbrodt, F. D. (2023). Big little lies: a compendium and simulation of p-hacking strategies. *R. Soc. Open Sci.* 10:220346. doi: 10.1098/rsos.220346

Stefani, L. A. (1998). Assessment in partnership with learners. *Assess. Eval. High. Educ.* 23, 339–350. doi: 10.1080/0260293980230402

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Front. Psychol.* 8:862. doi: 10.3389/fpsyg.2017.00862

Szucs, D. (2016). A tutorial on hunting statistical significance by chasing N. *Front. Psychol.* 7:1444. doi: 10.3389/fpsyg.2016.01444

TARG Meta-Research Group (2022). Statistics education in undergraduate psychology: a survey of UK curricula. *Collabra* 8:38037. doi: 10.1525/collabra.38037

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psychol.* 37, 1–2. doi: 10.1080/01973533.2015.1012991

Uher, J. (2018). Taxonomic models of individual differences: a guide to transdisciplinary approaches. *Philos. Trans. R. Soc. B Biol. Sci.* 373:20170171. doi: 10.1098/rstb.2017.0171

Uher, J. (2021a). Psychology's status as a science: peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integr. Psychol. Behav. Sci.* 55, 212–224. doi: 10.1007/s12124-020-09545-0

Uher, J. (2021b). Psychometrics is not measurement: unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.* 41, 58–84. doi: 10.1037/teo0000176

van 't Veer, A. E., and Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *J. Exp. Soc. Psychol.* 67, 2–12. doi: 10.1016/j.jesp.2016.03.004

Webb, M. A., and Tangney, J. P. (2022). Too good to be true: bots and bad data from mechanical Turk. *Perspect. Psychol. Sci.* 17456916221120027:174569162211200. doi: 10.1177/17456916221120027

Weiss, L. G., Saklofske, D. H., Coalson, D., and Raiford, S. E. (2010). *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. Cambridge, MA: Academic Press.

frontiers | Frontiers in Psychology

TYPE Conceptual Analysis
PUBLISHED 18 April 2024
DOI 10.3389/fpsyg.2024.1374330



OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
Valery Chirkov,
University of Saskatchewan, Canada
Nicolò Gaj,
Catholic University of the Sacred Heart, Italy

*CORRESPONDENCE
Kathleen L. Slaney
✉ klslaney@sfu.ca

RECEIVED 21 January 2024
ACCEPTED 02 April 2024
PUBLISHED 18 April 2024

CITATION
Slaney KL, Graham ME, Dhillon RS and Hohn RE (2024) Rhetoric of psychological measurement theory and practice. *Front. Psychol.* 15:1374330. doi: 10.3389/fpsyg.2024.1374330



# Rhetoric of psychological measurement theory and practice

Kathleen L. Slaney*, Megan E. Graham, Ruby S. Dhillon and Richard E. Hohn

Department of Psychology, Simon Fraser University, Burnaby, BC, Canada

Metascience scholars have long been concerned with tracking the use of rhetorical language in scientific discourse, oftentimes to analyze the legitimacy and validity of scientific claim-making. Psychology, however, has only recently become the explicit target of such metascientific scholarship, much of which has been in response to the recent crises surrounding replicability of quantitative research findings and questionable research practices. The focus of this paper is on the rhetoric of psychological measurement and validity scholarship, in both the theoretical and methodological and empirical literatures. We examine various discourse practices in published psychological measurement and validity literature, including: (a) clear instances of rhetoric (i.e., persuasion or performance); (b) common or rote expressions and tropes (e.g., perfunctory claims or declarations); (c) metaphors and other "literary" styles; and (d) ambiguous, confusing, or unjustifiable claims. The methodological approach we use is informed by a combination of conceptual analysis and exploratory grounded theory, the latter of which we used to identify relevant themes within the published psychological discourse. Examples of both constructive and useful or misleading and potentially harmful discourse practices will be given. Our objectives are both to contribute to the critical methodological literature on psychological measurement and connect metascience in psychology to broader interdisciplinary examinations of science discourse.

KEYWORDS

**psychological measurement, rhetoric, rhetoric of science, validation, metascience, methodological reform**

## Introduction

The theory and practice of psychological measurement has long been debated from numerous perspectives. Less represented in these topics, however, is the concern of how psychological researchers and measurement scholars *communicate* their findings and perspectives with respect to the construction, validation and use of measurement instruments in psychology. The focus of the present paper is, thus, on the *conceptual* arena of psychological measurement; that is, on the ways in which psychological researchers – both measurement and validity specialists and researchers using and reporting on psychological measurement tools – write about psychological measurement and validity, more generally.

First, we provide a brief overview of the rhetoric of science scholarship, including work examining the use of rhetoric in psychological research. We then summarize several different ways in which rhetoric appears in psychological measurement discourse. We describe several common forms of rhetoric and other styles of writing in psychological measurement and validity scholarship and provide examples from the broad theoretical psychological measurement and validity literatures. Our discussion is further supported by examples

collected from a sample of recently published research articles from a larger study we have been conducting on rhetoric of psychological science (Slaney and Wu, 2021; Slaney et al., 2024).

## Rhetoric of science

We begin by drawing a distinction between *discourse* and *rhetoric* and between *discourse analysis* and *analysis of rhetoric*. Whereas discourse extends to all forms of speech, writing, and communication, rhetoric is one of many possible features of discourse in which the speaker (writer, or communicator) intends to frame the message in such a way as to persuade or, at least, privilege a specific interpretation of the content at hand. Understood in this way, discourse analysis can be generally construed as the analysis of some form of speech, writing, or communication. The analysis of rhetoric pertains to analysis of forms of rhetorical discourse or rhetoric within a given discourse. The persuasive aspects of science discourse have long been recognized in philosophy of science circles (Overington, 1977). Science and technology studies scholars have also been concerned with tracking scientific discourse, oftentimes to analyze the legitimacy and validity of scientific claim-making (e.g., Zerbe, 2007). A subset of such scholarship has been concerned with rhetoric both as a feature of scientific discourse practice and a potential form of knowledge itself (Gross, 2006). Whereas the former contributes to the larger domain of metascience (i.e., serves as a way of understanding science and scientists; Gross, 2006), the latter is more epistemic in orientation (i.e., serves as a "way of knowing" itself).

*Rhetoric of science* is a subfield of this scholarship and is broadly defined as "the application of the resources of the rhetorical tradition to the texts, tables, and visuals of the sciences" (Gross, 2008, p. 1). It specifically concerns the forms of argumentation and persuasion that appear in scientific writing, including on philosophical, theoretical, and empirical topics relevant to science generally and within specific research domains. According to Kurzman (1988), rhetoric of science is central to the drawing of logical inferences (theoretical, empirical, statistical) by scientists. Further, Gaonkar (1993) states the "general aim of the [rhetoric of science] project is to show that the discursive practices of science, both internal and external, contain an unavoidable rhetorical component" (p. 267) and that "science is rhetorical all the way" (p. 268). Importantly, this should not be taken to suggest that science is *nothing more than* argument and attempted persuasion but, rather, that studying the rhetorical function and form of scientific discourses "has something important to contribute to our understanding of how science develops" (Ceccarelli, 2001, p. 177).

It is important to note that *metascience* has been viewed by some critical scholars as insufficient for dealing with deep-rooted conceptual problems within psychological science (e.g., Slaney, 2021; Malick and Rehmann-Sutter, 2022). We agree that metascience might leave little room for the examination of rhetoric and other forms of psychological science discourse if narrowly conceived as a domain of scholarship concerned only with whether the dominant methodology and methods of the natural sciences are being properly applied. However, here we advocate for a broader conception of metascience construed broadly as "science about science" or "research about research" and not restricted to either the natural sciences or to critiques of limited or faulty applications of quantitative methods. Framed in this way, metascience captures critical examinations of science discourse, connecting it to philosophy of science and science and technology studies scholarship, including rhetoric of science studies.[1]

## Rhetoric of psychological science

Psychology has only relatively recently become the explicit target of metascience scholarship on a broader scale but most of this has been in response to recent crises surrounding replicability of quantitative research findings and questionable research practices (QRPs) within the discipline (e.g., John et al., 2012; Open Science Collaboration, 2012, 2015; Lindsay, 2015). Despite work identifying common problematic discourse practices in the discipline (e.g., overly simplistic language; unclear, misleading or inaccurate content; and logical errors; Smedslund, 1991, 2015; Slaney and Racine, 2011, 2013; Lilienfeld et al., 2015; Slaney, 2017; Uher, 2022a,b), few studies have directly addressed the relevance of rhetoric of science scholarship for analyzing psychological science discourse or even recognized that psychological research has been both the target and a tool of rhetorical analysis (Carlston, 1987; Nelson et al., 1987; Bazerman, 2003).

Most of the work explicitly examining rhetoric in psychology has been done either by theoretical psychologists or critical scholars from other disciplines (e.g., science communication scholars; philosophers of science). The rhetorical aspects of the psychological research report have been the subject of some of the work of scholars external to the discipline. Bazerman (1987) traced the history of the "codification" of published research in psychology from stylesheets and supplements in the journal *Psychological Bulletin* through the first three revisions of the *American Psychological Association (APA) Publication Manual* (American Psychological Association, 1974, 1983).[2] Although the broad implementation of the *APA Publication Manual* facilitates communication and simplifies interpretation of research findings, Bazerman suggests the appearance of "epistemological neutrality" is "rhetorically naïve" and perpetuates a psychological research discourse that amounts to "incremental encyclopedism." In other words, the rigid APA publication format appears on the surface to merely "gather and report the facts" toward a progressively more and more complete description of behavior (Bazerman, 1987, p. 258, p. 273). For example, methods and results sections have become particularly technical and perfunctory, functioning more to protect researchers from claims of methodological error than to support innovative theory (Bazerman, 1987; John, 1992). In conforming to the highly accessible, yet excessively constraining, structure of the APA publication format, researchers do their best to appear to "tell it like it is" while at the same time putting their "best foot forward," both of which are clearly forms of rhetoric (i.e., attempted persuasion; Simons, 1993). Walsh and Billig

1 Uher (2023) uses "metatheory" to capture the philosophical and theoretical assumptions researchers hold about the phenomena they study. In the current work, because we focus on a set discourse *practices* within psychological science, we believe "metascience" better captures the kind of inquiry we are engaged in.

2 Four revisions have since been published, in 1994, 2001, 2009, and 2019, respectively.

(2014, p. 1682) asserted that the rhetorical style of the APA research report has become the "virtual *lingua franca*" of the discipline. Katzko (2002, p. 262) referred to it as an "institutionalized form of argumentation."

Carlston (1987) emphasized that, while it is true that the psychological research discourse is a legitimate *target* of rhetorical analysis, psychological research may also be a *tool* of such analysis because psychologists "study processes and phenomena that are central to language, stories, persuasion and other topics of rhetoric and hermeneutics" (p. 145). He asserted that many of the theoretical constructs at play in psychological discourse (e.g., "schema," "emotion," "memory," "motivation") are not just labels for the phenomena under study but, rather, are "summarizations of theories, histories, issues and arguments" (p. 147). Essex and Smythe (1999) echoed this notion and added that the reification of psychological constructs (i.e., treating them as concrete or objectively real) understood in terms of statistical correlations between scores on psychological measures is reinforced by a positivist legacy in psychological measurement theory and practice.

Rhetoric in psychological research discourse has also been examined from within the discipline (e.g., Danziger, 1990, 1996; Abelson, 1995; Morawski, 1996; Rose, 2011). Two of the most pervasive practices are what discourse analysts call nominalization and passivization (Billig, 1994, 2011, 2013). Nominalization is the use of nouns to express what are actually actions (e.g., "perception" instead of "to perceive") and passivization is researchers' use of passive phrasing in describing their own research activities (e.g., "A measure was administered" instead of "We administered a measure"; "Scores were obtained" instead of "We used the following scoring rule to form composite scores"). Billig (2011, 2013) argued such writing styles reify (i.e., create "fictional things") and "big up"[3] theoretical constructs by making them appear more noteworthy or intellectually rigorous. Such rhetoric gives the *appearance* of greater technical precision and objectivity and "depopulates" the texts of research discourse (i.e., of the *people* involved in the research; Billig, 1994). The problem with this is that although such writing styles may create more succinct discourse, when used to describe human actions, the sentences they produce tend to convey *less* information (e.g., about who is doing the actions and to whom; how the phenomenon of interest is being operationalized) than sentences using active verbs. Consequently, such terms can give the appearance of precision; yet the writer's meaning may remain inexplicit and ambiguous. Moreover, such writing styles reflect a prevalence of vague, abstract or unclear writing in psychological science (Billig, 2013; Kail, 2019).

Drawing from Billig's work, the first author of the current work has examined the rhetoric of psychological constructs, arguing that the heavy use in psychological research reports of passive voice and nominals in place of verb clauses has contributed to the reification of psychological constructs and the widespread ambiguity concerning the intended meanings of specific psychological constructs, as well as of the meaning of the term "construct" itself (Slaney and Garcia, 2015; Slaney, 2017). We argued such rhetoric provides a partial explanation for the pervasive practice in psychological discourse of confusing psychological *constructs* with the *phenomena* such constructs are intended to *represent*. Put another way, rhetoric partially explains why theoretical concepts (i.e., terms, conceptual models, theories) created by researchers are often confused with the phenomena those concepts are meant to describe. Where there are such ambiguities surrounding the ontological status of psychological constructs (i.e., what they *are*), it remains unclear what it would mean to "measure," "experimentally manipulate," "assess," "tap into," "investigate" or "validate" one, all of which are practices central to psychological measurement theory and validation.

In other work, we identified two areas in addition to the rhetoric of constructs in psychological research discourse: the *rhetoric of crisis* and the *rhetoric of methodology* (Slaney and Wu, 2021). The rhetoric of crisis refers to the more recent attention given to the "replication crisis" and a host of QRPs in psychology. The rhetoric of methodology represents a broader set of discourse practices, including rhetoric surrounding psychological measurement. The "quantitative imperative" identified by Michell (2003), according to which psychological attributes are presumed to have inherent quantitative structure and are therefore measurable, is one example (Michell, 2003). Another example is the pervasive "language of variables" which replaced the language of the "stimulus–response" unit in the latter half of the twentieth century to accommodate the then growing practice of building theory through the ongoing establishment of correlations among psychological measurements (Danziger, 1996; Toomela, 2008). A third example is the common practice of psychological researchers reporting that the measures used in their studies are "reliable and valid," often with no additional information or evidence about the psychometric properties of the measurement data from their studies (Weigert, 1970; Lilienfeld et al., 2015).

Additional critiques of conventional conceptions of and approaches to psychological measurement have identified other issues relevant to the present discussion. Tafreshi et al. (2016) argued that the quantitative imperative is one of several motivations for quantifying information in psychological research. Other motivations include the perceived need of ensuring objectivity, precision and rigor, reliance on statistical inference and adherence to both positivist and realist philosophies of science (Porter and Haggerty, 1997). In other work, the quantitative imperative has been addressed from a conceptual perspective, questioning the coherence of the very question of whether psychological attributes are measurable (see, for example, Maraun, 1998, 2021; Bennett and Hacker, 2022; Franz, 2022; Tafreshi, 2022; Tafreshi and Slaney, in press). Toomela (2008) argued that the implicit assumption that variables (i.e., data generated from the administration of psychological measures) directly represent the mental phenomena is based on faulty reasoning that there is a one-to-one correspondence between mental phenomena and behavior (i.e., measured variables). Lamiell (2013, p. 65) identified "statisticism" – the "virtually boundless trust of statistical concepts and methods to reveal" psychological laws – as fundamental way of thinking in contemporary psychological science. Uher (2022a,b) described several common conflations psychological and other social researchers make about measurement (e.g., data generation versus data analysis; quantity versus quality; measurement versus quantification). Bergner (2023) identified common scale construction practices based on confused concepts and flawed logic. It could

3 Smedslund's recent critique of "neuro-ornamentation" – the attempt to strengthen the impact of psychological study findings by inserting references to neuroscience – is another potent example of psychological researchers trying to "big up" the scientific relevance of their research (Smedslund, 2020).

be argued that these (and other) basic assumptions and practices of many psychological researchers are based more in a kind of perfunctory rhetoric than in scientific, theoretical or observational principles. Although they do not directly address the issue of rhetoric in psychological measurement literature, in a recent article, Flake and Fried (2020) identified an array of "questionable measurement practices" (QMPs), including everything from omissions of psychometric information to outright fraud and misrepresentation. One might contend that such "measurement flexibility," when used to misrepresent or steer interpretations of study findings in a particular direction is an abuse of "epistemic authority" (John, 1992) and a form of rhetoric that should be made transparent.

## The current study: rhetoric and other discourse practices in psychological measurement and validity discourse

In the current work, we aim to dig a little deeper into the discourse practices of psychological researchers, specifically those related to psychological measurement. Our primary objective is to provide concrete examples of some common ways of writing about the uses and validation of psychological measurement tools and identify their potential rhetorical features. We draw from two different literatures, the first being the broad theoretical and methodological literature on psychological measurement and validation, the second a sample of recently published research articles. We explore both constructive and useful or misleading and harmful uses of the discourse practices.

## Method and results

### Sample

To explore the rhetoric and other discourse practices relevant to measurement and validation in the empirical psychological research literature, we reviewed a sample of recently published research reports from a larger project we have been conducting on rhetoric of psychological science (Slaney and Wu, 2021; Slaney et al., 2024). The initial sample ($N = 40$) combined two samples (each with 20 articles) from separate studies, one of which focused on the uses of cognitive and causal metaphors (Subsample 1), the other on discourse related to null hypothesis statistical testing procedures (Subsample 2; see Table 1). Articles in both samples were randomly selected from larger article databases representing issues published in 2021 in APA journals across a range of subject categories[4] (~37 journals categorized as "Basic/experimental Psychology," "Developmental Psychology" and "Neuroscience & Cognition" for Subsample 1 and 50+ journals categorized as "Basic/experimental," "Clinical Psychology," "Developmental," "Forensic Psychology" and "Social Psychology & Social Processes" for Subsample 2). Due to overlap in the journals listed across the journal subject categories, we ensured that journals appeared only once. This created article populations of $N = 561$ and $N = 266$, respectively, for the first and second studies, from which we randomly sampled twenty articles from each. We included research reports on findings from quantitative data used in a single empirical study or on multiple studies reported in a single research report (i.e., by the same authors to address a set of hypotheses/research questions). We excluded editorials, commentaries, systematic reviews, non-English or strictly theoretical/methodological studies. One article from this sample was ultimately excluded, as the methods were deemed to be primarily qualitative with no use of quantitative measurement. Therefore, the final sample for the current study consisted of 39 articles.

### Procedure

Two research assistants independently reviewed and coded articles for a range of discourse practices including: (a) clear instances of rhetoric (i.e., persuasion or performance); (b) common or rote expressions and tropes (e.g., perfunctory claims or declarations); (c) metaphors and other "literary" styles; and (d) ambiguous, confusing, or unjustifiable claims. Coding categories were loosely defined *a priori*, though we left open the possibility of emergent themes.

Of the 39 articles, 20 were first reviewed and coded by both research assistants and the coding of the remaining 19 articles split between the two research assistants. Blocks of text were excerpted and then coded in terms of the categories described above. For those articles coded by both research assistants, overlapping excerpts were reconciled into a single entry in our textual database. We resolved discrepancies in coding through discussion with the entire research team and reflected finalized codes in the database. Though research assistants found multiple instances of a single code within a single article, the counts we report here of specific discourse practices capture the number of articles that contained at least one instance of a specific code. The final dataset was reviewed and vetted by the first author.

Before considering the results of this study, it is important to emphasize that our primary objective is not to make strong inferences strictly based on our sample about the prevalence of the discourse practices we have categorized herein. Rather, our main objective is to explore the conceptual landscape of validation and psychological measurement discourse practices – through *both* the theoretical and empirical literatures – to identify some of the ways in which psychological researchers use specific styles of writing to convey their understandings of measurement and validation tools, as well as the data generated from such tools. As such, the present study is better positioned as a conceptual analysis rather than as an empirical review of the theoretical and empirical psychological measurement and validation discourses at large. The results we present are meant to illuminate where such discourse practices are useful, benign or where they may be detrimental and potentially at odds with the intentions of psychological researchers.

### Results

#### Persuasive rhetoric of measurement

Michell (2003) argued the relevance and appropriateness of psychological measurement is almost universally *assumed* by

4 See https://www.apa.org/pubs/journals/browse?query=subject:Basic+%2f+Experimental+Psychology&type=journal

TABLE 1 Journals represented in each subsample.

| Sample | Journal |
|---|---|
| Subsample 1 | *Neuropsychology* |
| | *Group Dynamics: Theory, Research and Practice* |
| | *Journal of Consulting and Clinical Psychology* |
| | *Psychological Trauma: Theory, Research, Practice, and Policy* |
| | *Journal of Experimental Psychology: Animal Learning and Cognition* |
| | *Experimental and Clinical Psychopharmacology* |
| | *Journal of Diversity in Higher Education* |
| | *Sport, Exercise, and Performance Psychology* |
| | *Psychology of Violence* |
| | *Emotion* |
| | *Journal of Abnormal Psychology* |
| | *Psychology, Public Policy, and Law* |
| | *Journal of Experimental Psychology: General* |
| | *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* |
| | *American Journal of Orthopsychiatry* |
| | *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* |
| | *Journal of Family Psychology* |
| | *Psychological Assessment* |
| Subsample 2 | *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* |
| | *Psychology of Men & Masculinities* |
| | *Journal of Experimental Psychology: Human Perception and Performance* |
| | *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* |
| | *Neuropsychology* |
| | *Emotion* |
| | *Journal of Experimental Psychology: Applied* |
| | *Psychology of Aesthetics, Creativity, and the Arts* |
| | *Journal of Experimental Psychology: General* |
| | *Psychology and Aging* |
| | *Developmental Psychology* |
| | *Psychoanalytic Psychology* |
| | *Dreaming* |
| | *Experimental and Clinical Psychopharmacology* |
| | *Journal of Experimental Psychology: Learning, Memory, and Cognition* |

psychological researchers. Although this does not constitute an obvious attempt to persuade, that very few psychological researchers question the feasibility of psychological measurement could be seen as a form of implicit persuasion that pervades both theoretical and empirical psychological research discourses. Of course, there are more explicit forms of rhetoric surrounding psychological measurement validation. The very objective of validation research is to provide compelling evidence that a measure or measurement data are valid in one or more of the many senses that exist of psychometric validity. Such research clearly plays an important role in persuading readers and consumers of research that a given measurement tool meaningfully quantifies the putative trait it was designed to measure or assess. In fact, it is now very common in empirical research reports to include evidence for justifying the use of the measures used in the study at hand.

The importance of providing persuasive evidence for measurement tools is also reflected in methodological standards and guidelines of the discipline. For example, the American Psychological Association (APA) Publication Manual (American Psychological Association, 2020) specifies an array of *journal article reporting standards* (JARS),[5] including for reporting psychometric information concerning measurement data, the instruments used to generate these, and all

5 The JARS guidelines largely reflect those published in 2008 by the APA Publications and Communications Board Working Group on Journal Reporting Standards (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008), which were updated in 2018 (Appelbaum et al., 2018).

other relevant psychometric information. Although clear reporting standards are essential within any scientific discipline, it is important to acknowledge the potential drawbacks Bazerman (1987) and others have identified that accompany overly rigid codification of research reports. Perfunctory reporting of psychometric information is a poor replacement for clear *demonstration* that the measures used and measurements generated in research studies are appropriate for study objectives.

In our article sample,[6] we found examples of explicitly persuasive references to "important findings," "substantial links," "strong indicators," and "robust" measures (e.g., models, effects, etc.), and "rich and informative" theoretical models. Some of these claims were not supported directly with empirical evidence and in some cases even accompanied *weak* empirical evidence, counter to the descriptions of "strong" or "robust" findings. We also found less direct appeals to the importance of study findings, such as references to the production of "useful" knowledge, "novel findings," "advancing" knowledge in face of paucity of research or "gaps in the literature" and references to "confirming," "reaffirming," "reinforcing" expectations or findings from previously published research. Not surprisingly, most articles in our sample made as least one reference to "reliable" or "valid" measures or to the "reliability" or "validity" of the measures used in the study, over half of which (29 articles for "reliable"/"reliability" and 24 articles for "valid"/"validity") either reported no direct evidence or vaguely gestured to previously published psychometric evidence. Examples of each of these kinds of explicitly persuasive forms of rhetoric are given in Table 2.

## Common or rote expressions and tropes

As with methodology discourse practices generally, there are some expressions and turns of phrase that have become prevalent in psychological researchers' reporting of psychometric properties. As first illuminated by Weigert (1970), it is extremely common for psychological researchers to merely state that the measures used are "reliable and valid" or have "good," "acceptable" or "sufficient" reliability and validity, often with no definitions of or distinction made between these concepts or evidence provided for the putative reliability or validity of the measurements or measurement instruments in question. The use of such rote expressions presents numerous problems, including that reliability and validity are quite different psychometric properties and, in the case of validity, bear on multiple different aspects of measures and measurements and uses thereof; that both may be assessed with different metrics (depending on the nature of the scale of measurement); and that reliability is required for validity but not vice versa. Another problem is that ordinary and technical senses of reliability become conflated when references are made to reliable and valid *measures* as opposed to of *measurements* (i.e., data): To state that a *measure* (i.e., the measurement instrument itself) is reliable (i.e., dependable, suitable) is quite a different claim than to state that *measurements* (i.e., scores or data from administering the measure) have strong psychometric reliability (i.e., a low ratio of error variance to observed variance of scores on a random variable). Another example of rote-like reporting on psychological measurement is the common practice of cursorily reporting only traditional aspects of validity (i.e., content, criterion-oriented [predictive and concurrent] and construct), which fails to reflect the seven decades of validity theory and methodology since Cronbach and Meehl's seminal 1955 article (Cronbach and Meehl, 1955). In which validity was narrowly conceptualized in terms of these three broad types.

In our sample, phrases combining reliability and validity into a seemingly single psychometric property (i.e., "reliability and validity," "reliable and valid") did appear in the main body of some of the articles in our sample (see Table 2). The descriptor "good" was used often and to qualify everything from general reliability and validity or "psychometric/measurement properties" to specific kinds of validity (e.g., "model fit," "convergent") or reliability (e.g., "test–retest," "internal reliability," "stability," "agreement"). There appears to be at least some degree of rhetorical motivation for these appeals to "goodness," given that typically little elaboration was provided. Such underspecified claims appear to rhetorically stand in for any direct evidence of the psychometric properties of the measure being used to generate data for the study.

## Metaphors and other literary styles

The use of metaphors in scientific discourse is hardly rare and there have been many celebrated cases in the physical and life sciences (e.g., Bohr's "planetary" model of the hydrogen atom; evolutionary "tree" of life; DNA as a "twisted ladder"). Psychological measurement discourse also contains some commonly used metaphors, such a "tapping" "probing," and "emerging" in reference to putative fundamental factors or "constructs" said to "underlie" an observed correlation matrix of a set of item or subscale scores. Item-level scores are framed as "indicators" of "latent" factors, the latter of which are sometimes described as "driving" observed relations among item-level or subscale scores. Other common literary styles include the use of passive voice (e.g., "the measure *was administered* to…"; "…*was assessed by*…") and nominals in place of verb clauses (e.g., "…measure the *construct of* extraversion") of the kind Billig (2011) has identified. Both the uses of passive voice and nominalization of actions and activities of persons into traits presumed be "tapped" or "probed" by psychological measures constitute examples of depopulating texts, whereby the specific researchers making and acting upon decisions about the measurement tools used in their research become obscured. Such discourse styles serve a "rhetoric of scientificity" (Bourdieu, 1975) which is intended to give the impression that the research was conducted rigorously and objectively and, therefore, the findings can be trusted.

In our sample, each of the articles contained metaphors of one kind or another. The most common terms were "tap" (or "tapping") in relation to the phenomenon putatively measured or assessed and "reveal" (or "revealing") in reference to data or findings. We found that the terms "tap" and "reveal" were used to convey that measurement data had unveiled an underlying or latent realm. Across the sample, other common metaphors were "emerge/emerging" and "detect/detectable/detection." More unique metaphor use was exemplified by

6 Because we treated the text from our sample of articles as a qualitative source of data, we have indicated article numbers rather than formal citations in the results described, including directly excerpted text. Citations will be made available upon requests made to the first author.

10.3389/fpsyg.2024.1374330

TABLE 2 Article sample results.

| Sample article | Example |
|---|---|
| **Persuasive rhetoric of measurement** | |
| Explicitly persuasive forms of rhetoric | |
| Reported a "substantial link" between the independent and dependent variables where estimated effects were normatively small (i.e., r = 0.17 and d = 0.34). | Article 1 |
| Explicit reference to the importance of "objective measures," without elaboration of what constitutes objective in reference to the measure used. | Articles 6, 18, 33 |
| Stated the measure used in the study "has undergone rigorous evaluation and been found to perform well relative to similar measures," without reporting explicit psychometric evidence to justify. | Article 19 |
| Described instrument used in study as the "gold standard" for the assessment of the phenomenon without elaboration of why this marker of excellence was provided. | Article 36 |
| **Common or rote expressions and tropes** | |
| Vague gestures to previous research, validity, and reliability | |
| "Previous research has shown that…measures are more sensitive to [focal phenomenon]." | Article 2 |
| "Previous research finds the [measure]has adequate test–retest reliability." | Article 19 |
| "Previous research has demonstrated the validity of [the measure]." | Articles 28, 37 |
| Reported "reliability and validity" as a general property. | Articles 12, 13, 22, 27, 31 |
| **Metaphors and other literary styles** | |
| Metaphors | |
| Measure was described *"tap[ping]** children's ability to suppress a dominant response and undertake a subdominant response." | Article 5 |
| "The results *revealed** a significant three-way interaction between age group, condition, and perceived partner closeness." | Article 26 |
| References to "emerge" or "emerging" in relation to measured phenomena. | Articles 5, 12, 20, 21, 22, 27, 29 |
| References to "detect" or "detection" in relation to measured phenomena. | Articles 1, 2, 6, 8, 17, 19, 21, 22, 25, 26, 28, 30, 32, 33, 35, 38 |
| Use of "metaphorical story-telling" (Carlston, 1987). | Articles 16, 20 |
| **Use of passive voice** | |
| "The Structured Clinical Interview for the DSM–IV, nonpatient edition … *was administered** to assess for Axis I DSM–IV disorders." | Article 15 |
| *"Reward valuation ability** was assessed…" | Article 18 |
| **Misascribing actions or capacity** | |
| e.g., "the *measure** assessed" or "*items* access" as opposed to "We [the researchers] assessed … with the measure/items," "this *study** conceptualized…" instead of "We conceptualized…" | |
| A growing literature has explored…" instead of "A growing number of researchers have explored…" | Articles 3, 4, and 12 |
| **Confusing expressions, ambiguous, or unjustifiable claims** | |
| Construct validity | |
| "Such improvements in ADHD knowledge, use of behavioral strategies, and adaptive thinking skills, as measured by our study-specific measures, speak to their potential role as clinical change mechanisms, lending support to the construct validity of our *design**" | Article 3 |
| "[Cited authors] have provided evidence for the construct and criterion-related validity of this measure." | Article 31 |
| **Constructs** | |
| "As implicated in [cited study] meta-analysis, alliance is a *living*, * *evolving*, and *dynamic construct* that can be *perceived* and *reported* differently throughout the course of therapy." | Article 1 |
| Describe the construct of "functioning" as *representing** "a rather multifaceted construct, whose complexity may not have been captured by [the measure]." | Article 16 |
| Described the relationship between the focal construct and other constructs as follows: "anxiety, depression, and posttraumatic stress disorder (PTSD) are constructs that *display** significant overlap with alexithymia." | Article 18 |
| Generativity is a distinct construct *driven by** the *underlying desire* to contribute to the community and future generations through one's own legacy." | Article 34 |

*(Continued)*

TABLE 2 (Continued)

| Sample article | Example |
|---|---|
| **Missing Information** | |
| "It is beyond the scope of this article to report on all of the behavioral outcomes that were assessed in the current study but, in addition to measures of subjective response…" | Article 19 |
| **Hedging** | |
| "Various measurement approaches have been utilized in the field … Each of these measurement approaches has associated advantages as well as disadvantages and may capture distinct aspects of daily life." | Article 27 |
| **Other** | |
| Conflating ordinary and technical meanings of terms (e.g., reliable [*as in* dependable] measurement tools and measurements demonstrating high psychometric reliability). | Articles 1, 3, 5, 8, 17 and 30 |
| Conflating aggregate statistical findings with individual-level causal claims (e.g., "Previous research has demonstrated the validity of this manipulation, showing, for example, that social exclusion makes individuals more aggressive … and reduces prosocial behavior," and "Participants in the frustration condition further reported lower levels of satisfaction of the need for self-esteem"). | Article 28 |
| **Confusing statements** | |
| "[Cited article] reported that the [measure] can be applied in a four dimensional or unidimensional structure to collect data with good reliability and validity." | Article 13 |
| "…the experimental design could detect the presence/absence of the [measure] effect moderately well, but likely does not reliably detect small changes in the [measure] effect across conditions. To reliably detect a 15 ms change in the [measure] effect at roughly 80% power, for example, we estimate would require 100 participants per group." | Article 21 |

*Emphasis added.

"metaphorical storytelling" (Carlston, 1987), in which a concept or phenomenon is elaborated through a narrative style that relies on the use of metaphors. Examples of the use of these terms and discourse styles in our article sample are listed in Table 2.

We also found that the use of passive voice was ubiquitous in our article sample, appearing multiple times in every article (e.g., "was evaluated," "was assessed," "were measured," "were observed," "were obtained," etc.). It was also common, for example, to see such references to the administration of tests such as: "The Structured Clinical Interview for the DSM–IV, nonpatient edition … *was administered* to assess for Axis I DSM–IV disorders" (Article 15; emphasis added). This example is particularly noteworthy as the assessment tool in question is not a survey or trait measure, but a clinical interview, something that is inherently grounded in human interaction. To remove the interviewer from the "administration" of this test is indicative of the rhetoric of scientificity mentioned above.

In our sample, authors' use of nominals in place of verbs, as with the use of passive voice, was encountered in every article. This is not surprising, as it is virtually impossible to write efficiently without simplifying at least some verbal clauses with nominals (e.g., "perception" instead of "X perceived Y"), as Billig and discourse scholars have acknowledged. It has become so commonplace in social science writing that it is almost unnatural to describe human actions and capacities in verbal clauses.

Although not a literary device *per se*, it has become common in psychological discourse for writers to inappropriately ascribe to the subject of a sentence an action or capacity which could not, on logical grounds, be attributed to that subject (see examples in Table 2). Although such misattributions have become more common in contemporary discourse and often do not create too much confusion about what is being stated, they do contribute to the textual depopulating that Billig has identified as having a rhetorical aim.

## Confusing expressions, ambiguous, or unjustifiable claims

All forms of discourse at times contain unclear or confusing expressions; psychological scientific discourse is no exception. Although encountering the occasional ambiguous claim does not always create problems, science does not thrive in the face of pervasive ambiguity, and certainly not in unjustifiable statements. The discourse surrounding psychological "constructs" is one area where confusion, ambiguity and, in some cases, unjustifiable claims are commonly encountered.

Discussion of constructs pervades psychological research across theoretical, methodological and empirical domains. Yet, nowhere is there more ambiguity in the psychological measurement and validity discourse then with the "ever-evasive" construct concept (Slaney, 2017). Not only is the ontology of psychological constructs fuzzy, it is often difficult to discern what relationship constructs have to putative psychological "traits" and "mechanisms" ("qualities," "properties," "inferred entities," "processes," etc.); factors or "latent variables"; or with theoretical concepts, operational definitions, theories, theoretical statements, models or hypotheses (Maraun and Gabriel, 2013; Slaney, 2017). That is, constructs have been variously and confusingly characterized as *concepts* (e.g., theoretical constructs, hypotheses, models, theories), *objects of inquiry* (i.e., real but unobservable or only indirectly measurable theoretical *entities*, or features thereof) and, more generally, as the particular domain under study (e.g., "executive functioning," "prosociality," "attachment"). In fact, that psychological characteristics of persons are referred to as "traits," "mechanisms" and "processes" (and other such objectivist terminology) could be viewed as a form of rhetoric in presuming psychological attributes are just like physical traits, except that they are psychological in nature.

Although ambiguity is not itself an explicit form of rhetoric, if let unexamined it can carry rhetorical weight. For instance, in allowing constructs to be ontologically "fluid," some claims by researchers might appear stronger on the face of it than they really are. For

example, Colman (2006, p. 359) defines a (hypothetical) construct as "a conjectured entity, process, or event that is not observed directly but is assumed to explain an observable phenomenon." While this all sounds fine on the surface, it is unclear what it means for an "entity, process or event" to "explain" observable phenomenon. Although it has the ring of a precise scientific statement concerning the causal origins of the phenomenon under study, how the presence of causal structures and mechanisms could possibly be picked up by aggregate measurements is left unclear, at best. Similar ambiguities concerning the relationship between psychological constructs, observability and knowledge are prevalent in the discourse, as well as with other measurement-related concepts (e.g., "factor," "variable," "latent," "uni/multidimensional"; see, e.g., Green et al., 1977; Maraun and Gabriel, 2013; Slaney, 2017). As noted by Flake and Fried (2020), such "unjustified measurement flexibility" compromises the extent to which sound evidence about the measures used in a study can be provided which, in turn, casts doubt on the study findings overall.

In our sample, approximately half the articles referred to either of the terms "construct" or "construct validity." Construct validity was often claimed without direct appeal to psychometric evidence. For example, in some instances construct validity was presumed to be established through the common practice of simply invoking a previous single study. In one article, it was stated that "[s]uch improvements in ADHD knowledge, use of behavioral strategies, and adaptive thinking skills, as measured by our study-specific measures, speak to their potential role as clinical change mechanisms, lending support to the construct validity of our *design*" (Article 3; emphasis added). The references to both "clinical change mechanisms" and construct validity are vague, leaving unclear what is meant by the terms themselves, what the "construct" that has been validated is and how the results evidence the putative validity of said construct.

In terms of constructs themselves, authors from our sample referred to these without providing much if any indication of the specific natures of the constructs at hand. Several examples are listed in Table 2. Taking these examples together, it is difficult to determine the nature of psychological constructs such that they can be "driven by underlying" emotional states and considered to be "living" and "evolving," but also to "represent" putative traits (attributes, etc.) and "display" relationships with other constructs.

We found other confusing or ambiguous forms of writing in our sample. These include reference to missing information and hedging. Additional examples include conflating ordinary and technical meanings of psychological concepts as well as conflating aggregate statistical findings with individual-level causal claims. We also found a small number of completely unclear or confusing statements. Examples of each of these kinds of confusing and/or ambiguous claims can be found in Table 2.

# Discussion

## What's the problem with a little rhetoric?

### Constructive versus destructive rhetoric

It is important to note that rhetoric of science scholars are not united in how they frame rhetoric in science discourse or whether they view it as useful and essential, harmful and misleading, or inevitable or avoidable. Haack (2007, pp. 217–223) draws an important distinction between "reasonable" and "radical" rhetoric of science and between "modes of communication that promote the epistemologically desirable correlation, and those that impede it." She contrasts between two very different scenarios, one in which a scientific claim is accepted because clear and strong evidence is clearly communicated and the other in which a scientific claim comes to be accepted in the *absence* of good evidence because it is promoted by means of "emotive language, snazzy metaphors,… glossy photographs, melodramatic press conferences, etc." (p. 223). Whereas Haack describes the first scenario as legitimately persuasive, she views the second as "strictly rhetorical." Simons (1993) echoes something similar, noting that rhetorical argumentation does not necessarily make for bad argumentation; however, the slope from rhetoric to fraud may be slippery (Simons, 1993). More optimistically, Carlston (1987) characterizes an intertwining relationship between rhetoric and empirical science, wherein "empirical efforts complement but do not replace rhetorical practices, and rhetorical analysis illuminates but does not invalidate empirical pursuits," and both are legitimate tools for accumulating "useful understandings and knowledge" (p. 156).

For example, on the use of scientific metaphors as one potential rhetorical strategy, Haack (2007) concedes that although they "oil the wheels of communication" and can be a source of new and important avenues of inquiry, "their worth…depends on the fruitfulness of the intellectual territory to which these avenues lead" (p. 227). Further, Haack notes, a given scientific metaphor may lead scientists in different directions, some better, some worse. As Nagel (1961; as cited in Carlston, 1987) warned over six decades ago, the use of scientific metaphors can be detrimental if the limits of their uses are not properly acknowledged and attended to.

It is fair to ask why scientists would not genuinely wish to persuade readers and consumers to accept research findings they believe are based on strong scientific practice. We agree with Haack that it would be quite counter-intuitive for psychological or any other researchers to avoid making persuasive claims that their research findings are both valid and important. At the same time, it is not always fully clear or agreed upon as to what constitutes "strong" or "good" evidence. Simply claiming strong or good evidence is questionable rhetoric. Moreover, there is no necessary connection between radical (poor) rhetoric and bad (weak) evidence: One can use radical rhetoric in reference to valid and strong evidence and reasonable rhetoric in reference to poor evidence.[7] On the basis of the current sample of psychological research reports, we see that although some uses of rhetorical writing are relatively harmless (e.g., some nominalization, especially when its use is explicitly justified as descriptive efficiency) or even useful (e.g., metaphorical "story-telling" to clarify a concept), others create ambiguity, at least, and outright confusion, at worst. For example, sometimes using "variable," "factor," "construct," etc. interchangeably is harmless, as the intended meanings of these terms in some contexts need not be precise (e.g., in highly general references to the phenomenon under study); however, in other instances, conflating these terms can be truly confusing, such as when constructs are portrayed as theoretical (explanatory) models and *at the same time* the putative trait measured by a given instrument.

7 We thank an anonymous reviewer for highlighting this.

Clearly a construct cannot both be a theory and that which is the subject of the theory. Moreover, reifying aspects of psychological functioning through nominalization and other styles of discourse (e.g., "trait" terminology) can also affirm naïve naturalist and realist views on the nature of psychological reality, thus obscuring important conceptual connections between ordinary and scientific senses of psychological concepts (Danziger, 1990; Brock, 2015; Slaney, 2017; Tafreshi, 2022; Tafreshi and Slaney, in press).

### Why is studying rhetoric and other discourse practices in psychological measurement scholarship important?

Of course, the answer to this question depends on who you ask, as even rhetoricians are divided on the question of where rhetorical analysis fits within the grand scheme of science (Simons, 1993). As noted at the beginning of the paper, we view examining rhetorical and other discourse practices as an important part of metascience, a primary aim of which is to improve science through better understanding *of* science (Ceccarelli, 2001), or of a given discipline or area of study (Overington, 1977) as it evolves within current social contexts. As such, it constitutes a part of recent movements within the discipline to acknowledge and address fundamental problems with psychological research (e.g., replication crisis; fraud; identification of QRPs, QMPs, etc.) and, in so doing, improve psychological science (e.g., Society for the Improvement of Psychological Science [SIPS]).[8] We emphasize psychological measurement discourse not because it is unique in involving rhetorical features but because psychological measurement – even if not always explicitly acknowledged – provides the foundation for psychological research methods, more broadly. That is, a prevalence of questionable *measurement* practices "pose a serious threat to cumulative psychological science" and, yet, have received much less scrutiny and attention than failures of replication and other QRPs (Flake and Fried, 2020, p. 457), neither of which can be fully understood in the face of potentially widespread invalidity of the psychological measurement tools that generate the data which are the inputs for other psychological research methods.

It is also important to acknowledge that rhetoric and other discourse practices that might misrepresent the phenomena under study or otherwise create ambiguity or confusion occur neither in isolation nor in a vacuum. Most psychological research reports, including those in our sample, have been subject to peer and editorial review prior to publication.[9] Yet, problematic discourse practices, such as those we have identified, manage to make it past the peer-review and editorial filters. This signals that the use of confusing or unclear language (rhetorical or otherwise) in psychological research discourse is a systemic problem, not to be blamed just on individual researchers. As with other QRPs that threaten the integrity of psychological research, a response is needed to address the *questionable discourse practices* in psychology that have been illuminated here and elsewhere. How researchers frame their theoretical positions, methods choices, the data that arises from their implementation, and the interpretations they make of findings should be, we argue, an essential part of the discussion about QMPs and QRPs. The upside is that illuminating the detrimental effects of such practices can, if taken seriously, be rectified by broad implementation of training in such areas as philosophy of science, metatheory, and scientific writing for psychology (Billig, 2013; Slaney, 2017; Kail, 2019; Uher, 2023). We believe that exposing pervasive hidden assumptions researchers take into their research can influence how reflective researchers (and, by extension, the discipline) will be regarding the relevant subject matters they are concerned with. We see the current work, and that of other critical methods scholars, as making important contributions to current discussions about methodological crisis and reform.

8 https://improvingpsych.org/mission/

9 We thank an anonymous reviewer for raising this point.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent was not required in accordance with the national legislation and the institutional requirements.

## Author contributions

KS: Validation, Formal Analysis, Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization. MG: Validation, Writing – review & editing, Writing – original draft, Supervision, Investigation, Formal Analysis, Data curation. RD: Writing – review & editing, Writing – original draft, Investigation, Data curation. RH: Writing – review & editing, Writing – original draft.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

# References


Abelson, R. P. (1995). *Statistics as principled argument*. New York: Psychology Press.

American Psychological Association. (1974). *Publication manual of the American Psychological Association* (*2nd Edn*). Washington, DC: American Psychological Association.

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (*3rd Edn*). Washington, DC: American Psychological Association.

American Psychological Association. (2020). *Publication manual of the American Psychological Association 2020: the official guide to APA style* (7th ed.). American Psychological Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: why do we need them? What might they be? *Am. Psychol.* 63, 839–851. doi: 10.1037/0003-066X.63.9.839

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., and Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. *Am. Psychol.* 73, 3–25. doi: 10.1037/amp0000191

Bazerman, C. (2003). *Shaping written knowledge: The genre and activity of the experimental article in science*. University of Wisconsin Press: Madison.

Bazerman, C. (1987). "Codifying the social scientific style" in *Rhetoric of the human sciences: Language and argument in scholarship and public affairs*. ed. J. S. Nelson. *Revised* ed (Wisconsin, USA: University of Wisconsin Press), 257–277.

Bennett, M. R., and Hacker, P. M. S. (2022). *Philosophical foundations of neuroscience. 2nd Edn*. Hoboken, NJ: John Wiley & Sons.

Bergner, R. M. (2023). Conceptual misunderstandings in mainstream scale construction: suggestions for a better approach to concepts. *Theory Psychol.* 33, 701–716. doi: 10.1177/09593543231177696

Billig, M. (1994). Repopulating the depopulated pages of social psychology. *Theory Psychol.* 4, 307–335. doi: 10.1177/0959354394043001

Billig, M. (2011). Writing social psychology: fictional things and unpopulated texts. *Br. J. Soc. Psychol.* 50, 4–20. doi: 10.1111/j.2044-8309.2010.02003.x

Billig, M. (2013). *Learn to write badly: How to succeed in the social sciences*. Cambridge, NY: Cambridge University Press.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information*, 14, 19–47.

Brock, A. (2015). "The history of psychological objects" in *The Wiley handbook of theoretical and philosophical psychology: Methods, approaches, and new directions for social sciences*. eds. J. Martin, J. Sugarman and K. L. Slaney (Hoboken, NJ: Wiley Blackwell), 151–165.

Carlston, D. E. (1987). "Turning psychology on itself" in *The rhetoric of the human sciences: Language, and argument in scholarship and public affairs*. eds. J. S. Nelson, A. Megill and D. N. McCloskey (Madison, Wis: University of Wisconsin Press), 145–162.

Ceccarelli, L. (2001). *Shaping science with rhetoric: The cases of Dobzhansky, Schrödinger, and Wilson*. Chicago, IL: University of Chicago Press.

Colman, A. M. (2006). *A dictionary of psychology*. Oxford: Oxford University Press.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity and psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge: Cambridge University Press.

Danziger, K. (1996). "The practice of psychological discourse" in *The historical dimensions of psychological discourse*. eds. C. F. Graumann and K. J. Gergen (Cambridge: Cambridge University Press), 17–35.

Essex, C., and Smythe, W. E. (1999). Between numbers and notions: a critique of psychological measurement. *Theory Psychol.* 9, 739–767. doi: 10.1177/0959354399096002

Flake, J. K., and Fried, E. I. (2020). Measurement schmeasurement: questionable measurement practices and how to avoid them. *Psychol. Sci.* 3, 456–465. doi: 10.1177/2515245920952393

Franz, D. J. (2022). "Are psychological attributes quantitative?" is not an empirical question: conceptual confusions in the measurement debate. *Theory Psychol.* 32, 131–150. doi: 10.1177/09593543211045340

Gaonkar, D. P. (1993). The idea of rhetoric in the rhetoric of science. *South Commun. J.* 58, 258–295. doi: 10.1080/10417949309372909

Green, S. B., Lissitz, R. W., and Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of unidimensionality. *Educ. Psychol. Meas.* 37, 827–838. doi: 10.1177/001316447703700403

Gross, A. G. (2006). *Starring the text: The place of rhetoric in science studies*. Carbondale, IL: Southern Illinois University Press.

Gross, A. G. (2008). "Rhetoric of science" in *The international encyclopedia of communication*. ed. W. Donsbach (Carbondale, IL: Southern Illinois University Press)

Haack, S. (2007). *Defending science—Within reason: Between scientism and cynicism*. Amherst, NY: Prometheus Books.

John, I. D. (1992). Statistics as rhetoric in psychology. *Aust. Psychol.* 27, 144–149. doi: 10.1080/00050069208257601

John, L., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Kail, R. V. (2019). *Scientific writing for psychology: Lessons in clarity and style*. Thousand Oaks, CA: SAGE.

Katzko, M. W. (2002). The rhetoric of psychological research and the problem of unification in psychology. *Am. Psychol.* 57, 262–270. doi: 10.1037/0003-066X.57.4.262

Kurzman, C. (1988). The rhetoric of science: strategies for logical leaping. *Berkeley J. Sociol.* 33, 131–158.

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology*, 31, 65–71.

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., and Waldman, I. D. (2015). Fifty psychological and psychiatric terms to avoid: a list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Front. Psychol.* 6:1100. doi: 10.3389/fpsyg.2015.01100

Lindsay, D. S. (2015). Replication in psychological science. *Psychol Sci* 26, 1827–1832. doi: 10.1177/0956797615616374

Malick, L., and Rehmann-Sutter, C. (2022). Metascience is not enough – a plea for psychological humanities in the wake of the replication crisis. *Rev. Gen. Psychol.* 26, 261–273. doi: 10.1177/10892680221083876

Maraun, M. D. (1998). Measurement as a normative practice: implications of Wittgenstein's philosophy for measurement in psychology. *Theory Psychol.* 8, 435–461. doi: 10.1177/0959354398084001

Maraun, M. D. (2021). Language and the issue of psychological measurement. *J. Theor. Philos. Psychol.* 41, 208–212. doi: 10.1037/teo0000188

Maraun, M. D., and Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas Psychol.* 31, 32–42. doi: 10.1016/j.newideapsych.2011.02.006

Michell, J. (2003). The quantitative imperative: positivism, naïve realism and the place of qualitative methods in psychology. *Theory Psychol.* 13, 5–31. doi: 10.1177/0959354303013001758

Morawski, J. (1996). "Principles of selves: the rhetoric of introductory textbooks in American psychology" in *Historical dimensions of psychological discourse*. eds. C. F. Graumann and K. J. Gergen (Cambridge: Cambridge University Press), 145–162.

Nelson, J. S., Megill, A., and McCloskey, D. N. (1987). "Rhetoric of inquiry" in *The rhetoric of the human sciences: Language and argument in scholarship and public affairs*. eds. J. S. Nelson, A. Megill and D. N. McCloskey (Madison, Wis: University of Wisconsin Press), 3–18.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* 7, 657–660. doi: 10.1177/1745691612462588

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac 4716. doi: 10.1126/science.aac4716

Overington, M. A. (1977). The scientific community as audience: toward a rhetorical analysis of science. *Philosophy & Rhetoric* 10, 143–164.

Porter, T. M., and Haggerty, K. D. (1997). Trust in numbers: the pursuit of objectivity in science & public life. *Can. J. Sociol.* 22:279.

Rose, A. C. (2011). The invention of uncertainty in American psychology: intellectual conflict and rhetorical resolution, 1890-1930. *Hist. Psychol.* 14, 356–382. doi: 10.1037/a0023295

Simons, H. W. (1993). "The rhetoric of the scientific research report: 'drug-pushing' in a medical journal article" in *The recovery of rhetoric: Persuasive discourse and disciplinarily in the human sciences*. eds. R. H. Roberts and J. M. M. Good (London, UK: Bristol Classical Press), 148–163.

Slaney, K. L. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. London: Palgrave Macmillan.

Slaney, K. L., and Garcia, D. A. (2015). Constructing psychological objects: the rhetoric of constructs. *J. Theor. Philos. Psychol.* 35, 244–259. doi: 10.1037/teo0000025

Slaney, K. L., Graham, M. E., and Dhillon, R. (2024). Rhetoric of Statisical Significance. Manuscript in preparation.

Slaney, K. L. (2021). "Is there a waning appetite for critical methodology in mainstream scientific psychology?" in *Problematic research practices and inertia in scientific psychology: History, sources, and recommended solutions*. eds. J. T. Lamiell and K. L. Slaney (New York: Routledge), 86–101.

Slaney, K. L., and Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas Psychol.* 31, 4–12. doi: 10.1016/j.newideapsych.2011.02.003

Slaney, K. L., and Racine, T. R. (2011). On the ambiguity of concept use in psychology: is the concept 'concept' a useful concept? *J. Theor. Philos. Psychol.* 31, 73–89. doi: 10.1037/a0022077

Slaney, K. L., and Wu, C. A. (2021). "Metaphors, idioms, and Clichés: the rhetoric of objectivity in psychological science discourse" in *Routledge international handbook of theoretical and philosophical psychology*. eds. B. Slife, S. Yanchar and F. Richardson (New York, NY: Routledge), 453–472.

Smedslund, J. (1991). The Pseudoempirical in psychology and the case for psychologic. *Psychol. Inq.* 2, 325–338. doi: 10.1207/s15327965pli0204_1

Smedslund, J. (2015). Why psychology cannot be an empirical science. *Integr. Psychol. Behav. Sci.* 50, 185–195. doi: 10.1007/s12124-015-9339-x

Smedslund, J. (2020). "Neuro-ornamentation in psychological research" in *Respect for thought: Jan Smedslund's legacy for psychology*. eds. T. G. Lindstad, E. Stänicke and J. Valsiner (Cham, Switzerland: Springer)

Tafreshi, D. (2022). Sense and nonsense in psychological measurement: a case of problem and method passing one another by. *Theory Psychol.* 32, 158–163. doi: 10.1177/09593543211049371

Tafreshi, D., and Slaney, K. L. (in press). Science or not, conceptual problems remain: Seeking conceptual clarity around "psychology as a science" debates. *Theory Psychol.*

Tafreshi, D., Slaney, K. L., and Neufeld, S. D. (2016). Quantification in psychology: critical analysis of an unreflective practice. *J. Theor. Philos. Psychol.* 36, 233–249. doi: 10.1037/teo0000048

Toomela, A. (2008). Variables in psychology: a critique of quantitative psychology. *Integr. Psychol. Behav. Sci.* 42, 245–265. doi: 10.1007/s12124-008-9059-6

Uher, J. (2022a). Functions of units, scales and quantitative data: fundamental differences in numerical traceability between sciences. *Qual. Quant.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: a rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Soc. Personal. Psychol. Compass* 17:e12740. doi: 10.1111/spc3.12740

Walsh, R. T., and Billig, M. (2014). "Rhetoric" in *Encyclopedia of critical psychology: Springer reference*. ed. T. Teo (Berlin: Springer), 1677–1682.

Weigert, A. (1970). The immoral rhetoric of scientific sociology. *Am. Sociol.* 5, 111–119.

Zerbe, M. J. (2007). *Composition and the rhetoric of science: Engaging the dominant discourse*. Carbondale: Southern Illinois University Press.

frontiers | Frontiers in Psychology

TYPE Conceptual Analysis
PUBLISHED 21 May 2024
DOI 10.3389/fpsyg.2024.1354392



OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
David Torres Irribarra,
Pontificia Universidad Católica de Chile, Chile
Leslie Pendrill,
Research Institutes of Sweden (RISE), Sweden

*CORRESPONDENCE
Michele Luchetti
michele.luchetti@uni-bielefeld.de

†PRESENT ADDRESS
Michele Luchetti,
Department of Philosophy, University of Bielefeld, Bielefeld, Germany

RECEIVED 12 December 2023
ACCEPTED 29 April 2024
PUBLISHED 21 May 2024

CITATION
Luchetti M (2024) Epistemic circularity and measurement validity in quantitative psychology: insights from Fechner's psychophysics.
*Front. Psychol.* 15:1354392.
doi: 10.3389/fpsyg.2024.1354392



# Epistemic circularity and measurement validity in quantitative psychology: insights from Fechner's psychophysics

Michele Luchetti*†

Max Planck Institute for the History of Science, Berlin, Germany

The validity of psychological measurement is crucially connected to a peculiar form of epistemic circularity. This circularity can be a threat when there are no independent ways to assess whether a certain procedure is actually measuring the intended target of measurement. This paper focuses on how Fechner addressed the measurement circularity that emerged in his psychophysical research. First, I show that Fechner's approach to the problem of circular measurement involved a core idealizing assumption of a shared human physiology. Second, I assess Fechner's approach to this issue against the backdrop of his own epistemology of measurement and the measurement context of his time. Third, I claim that, from a coherentist and historically-situated perspective, Fechner's quantification can be regarded as a first successful step of a longer-term quantification process. To conclude, I draw from these insights some general epistemological reflections that are relevant to current quantitative psychology.

KEYWORDS

**quantification, Fechner, psychophysics, psychology, measurement, validity**

## 1 Introduction

The historical development of psychology as a science has been closely intertwined with the reflection on what counts as a psychological measurement. Several innovative developments in measurement theory over the twentieth century have directly stemmed from the work of psychologists and psychometricians, such as L. L. Thurstone, D. T. Campbell, S. S. Stevens, and R. D. Luce. Still today, the meaning and validity of psychological measurements represents a central concern for methodologists of quantitative psychology, to the point that some critics have questioned the very legitimacy of psychology as a quantitative discipline (e.g., Michell, 1997, 1999, 2008, 2012). Indeed, despite the use of quantitative methods is widely established in several areas of psychology, foundational conceptual and epistemological questions concerning the quantitative status of psychological entities and the use of quantitative methods in psychology are far from being settled.

In the period spanning from the origins of psychology as a quantitative science, in the second half of nineteenth century, up to the beginning of the twentieth century, the effort toward quantification concerned mainly two areas: psychophysics and mental testing (cf. Hornstein, 1988). Both areas were faced with the challenge of quantitatively representing characteristics, such as sensation and intelligence, which could not be directly observed.

The impossibility to measure these characteristics directly,[1] opened fundamental questions relative to what kind of measurement proxies could be considered as informative about the characteristic of interest and on what epistemological basis. In this paper, I focus on the early history of one of these enterprises, viz., Fechner's psychophysical project of quantifying sensation. Fechner's work can be regarded as a methodological laboratory for quantitative psychology, in that he engaged very early on with foundational measurement problems which became central to both psychophysics and psychology in general.

Fechner's philosophy of science and his theory of measurement were quite sophisticated. Since they have been extensively analyzed elsewhere, providing an overarching account of either of the two is beyond the scope of this contribution.[2] Instead, I will put one specific aspect of Fechner's approach to measurement at the center of my analysis, that is, his way of addressing the problem of epistemic circularity in measurement. This is the issue of how scientists justify their belief that certain measurement procedures identify a quantity or property of interest in the absence of independent methods to assess these procedures. This issue was a central concern for the success of his psychophysical project, as it is to current discussions on the validity of psychological measurements. Therefore, examining Fechner's work can, in my view, provide us with valuable insights to reflect on how to frame and address this problem from an epistemological perspective.

Before turning to my analysis, some important considerations are in order. Fechner's psychophysical project aimed at providing a quantification of experience, which he operationalized as the intensity of the internal sensations produced by physical stimuli. Therefore, it may be asked to what extent we can draw a fruitful comparison between epistemological issues concerning, respectively, the measurement of sensations of physical stimuli and the measurement of more complex psychological properties, such as intelligence or memory. The possibility of such an inferential step is connected to questions concerning the nature of psychological kinds and the definition of psychological constructs. On the one hand, psychological kinds seem to be quite different from other natural or scientific kinds, in that they are very multifaceted, their causal interactions produce effects that vary highly depending on context, and they undergo constant change. Therefore, psychological constructs seem to be better characterized as concepts representing clusters, or networks, or features of phenomena, rather than as monolithic attributes (Feest, 2017, 2022a). In addition, psychological constructs should reflect the changeability of psychological phenomena and be changeable themselves (Hanfstingl, 2019). Indeed, these features represent some of the central challenges to quantification in psychology (Uher, 2020, 2021a).

Fechner's challenge was that of finding ways to express "the amount of a psychological attribute with respect to something that was related to it in a spatio-temporal sense" (Briggs, 2021: p. 32), that is, a way to relate our internal experience, viz. sensation, to an external perceptible standard. In his view, as I will discuss, this could be tackled in the same way as for physical measurement, since he rejected any reason to restrict measurement to physical properties. However, we can see, even intuitively, that constructs like intelligence or memory are more complex and multi-dimensional than sensations. This is because these constructs refer to psychical performances which emerge through the joint manifestation of several different abilities, such as verbal knowledge, reading comprehension, etc (Toomela, 2008). Most importantly, the methods by which we can access these different phenomena vary, depending on the nature of the phenomena themselves. The response to physical stimuli can be studied through *extraquestive* methods, based on the possibility of establishing a shared perception of a physical phenomenon, both internal and external to individuals' bodies (Uher, 2019). However, these methods are not available for the study of internal psychic phenomena, that can be perceived only by each individual. These must be studied through *intraquestive* methods, which necessarily rely on language and interpretation by both the individuals acting as measurement instruments (the raters) and the scientists.[3]

In sum, features related to the multi-dimensionality and complexity of the psychological subject matter worsen the impact of certain general issues, such as those related to the possibility of experimental control (Trendler, 2009; Wajnerman-Paz and Rojas-Líbano, 2022).[4] On the other hand, psychological measurement presents specific conceptual, methodological, and epistemological challenges, compared to sensory measurement, due to both the peculiar nature of the phenomena under investigation and the limitations characterizing the appropriate methods currently available to study them.[5] Nonetheless, this does not mean that some fundamental issues characterize both sensory measurement and the measurement of more complex psychological phenomena. Indeed, the problem of epistemic circularity in measurement represents an issue that, despite manifesting itself in different ways and with different intensities, concerned both Fechner's sensory measurement and contemporary quantitative psychology. Given this level of abstraction, my insights on Fechner's approach to this problem will not translate into methodological maxims directly

1 The distinction between direct and indirect measurement methods is neither univocal nor uncontroversial. According to certain measurement traditions, this distinction collapses even in the case of intuitively direct physical measurements, e.g.: " [. . .] all measurements are indirect in one sense or another. Not even simple physical measurements are direct, as the philosophically naïve individual is likely to maintain. The physical weight of an object is customarily determined by watching a pointer on a scale. No one could truthfully say that he 'saw' the weight." (Guilford, 1936: p. 3).

2 Heidelberger (2004) offers a comprehensive account of Fechner's philosophy, including his philosophy of science and his theory of measurement. Briggs (2021) focuses more specifically on Fechner's meta-perspective on measurement and several technical aspects of great epistemological relevance.

3 For instance, conceptual errors involved by naïve uses of verbal items as measurement scales raise concerns that are distinctive to psychological measurement. Cf., for instance, Lundmann and Villadsen (2016), Smedslund (2016), and Uher (2022).

4 Another example of difference in challenges between psychophysics and other areas of psychology comes from the phenomenon of reactivity, i.e., the fact that humans may respond to their awareness of being studied, which manifests itself differently in different psychological contexts of research (e.g., Orne, 1962; Feest, 2022b). As such, reactivity is plausibly lower in the context of measuring sensory reactions than when the measurement process involves more complex language-based abilities, as in the case of higher-order psychological properties. On the pervasiveness of reactivity in the human sciences see Marchionni et al. (2024) and references therein.

5 See, for instance, Uher (2021b) for a comprehensive analysis of the conceptual and epistemological challenges to contemporary psychological measurement.

applicable to current psychological measurement. Rather, it will provide some broad epistemological considerations relative to two specific aspects: (1) the role of implicit untested measurement assumptions; (2) what counts as successful measurement and how it impacts general epistemic categories like validity and objectivity.

In section "2 Epistemic circularity and psychological measurement," I will introduce the epistemic problem of circular measurement, focusing specifically on psychological measurement and its challenges. In section "3 Fechner's psychophysics and the making of sensation as a quantity," I will first present Fechner's psychophysical research program in general and then zoom in on his approach to the problem of circular measurement. In section "4 Epistemological insights from Fechner's quantification of sensation," I will develop the main argument. First, I will focus on some relevant objections to Fechner's quantification of sensation raised by both his contemporaries and more recent commentators. Then, I will analyze Fechner's approach to measurement circularity and I will discuss it against the backdrop of Fechner's broader epistemology of measurement. Finally, I will reconsider Fechner's contribution vis-à-vis the subsequent history of psychophysical measurement. Section "5 The relevance of Fechner to current methodology of psychological measurement" will conclude by offering some insights on how the present work is relevant to contemporary quantitative psychology.

## 2 Epistemic circularity and psychological measurement

From an epistemological point of view, the problem of what counts as a good, reliable, or accurate measurement is connected with the problem of how to appropriately identify the target of measurement, that is, which concepts or constructs appropriately represent the measurand (Tal, 2019). These issues have indeed been a central focus of methodological debates in psychological measurement. However, I will first present how they have been tackled in recent philosophical and metrological literature as issues that concern measurement across the sciences.

Measurement procedures are often described as concrete interactions between one or more epistemic subjects (observers and/or test subjects), a material apparatus, and some phenomenon occurring in an environment. Examples of this are when we observe the mercury dilate in the column of a thermometer hanging on the wall or when a person responds to a standardized item on a personality test questionnaire. In the first case, the physical process itself that takes place during the measurement interaction can also be used to *represent* a certain relationship between quantities, as when we read a measurement of temperature out of an indication of the length reached by the mercury in the thermometer column. In the second case, the measurement interaction presupposes a certain representational relationship between measured items and certain target properties, as when scores attributed to individual responses of a personality test questionnaire are taken to be informative about a certain personality trait.

The fact that measurement has both a material and a representational dimension is central to an epistemic conundrum,

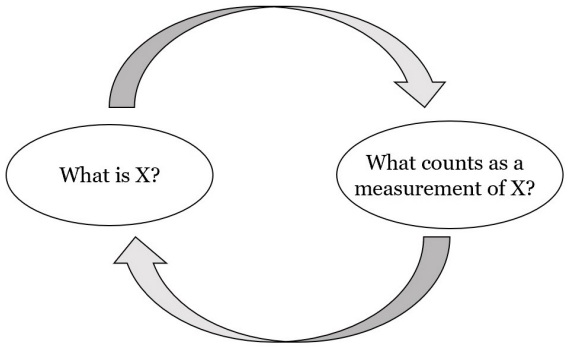



FIGURE 1
A graphic representation of the general problem of epistemic circularity in measurement. The arrows represent the direction of epistemic access.

namely, the problem of circular measurement.[6] This is the issue of how scientists justify their belief that certain measurement procedures identify the quantity or characteristic of interest in the absence of independent methods to assess these procedures. In the case of measuring a physical quantity, for example, we often infer its value from the values of other quantities, as when we infer measurement outcomes of temperature from indications of length of a thermometer column. This inference is based on knowledge of the empirical relationship between the quantities of temperature and length in a specific physical interaction. However, knowledge of this relationship is itself a scientific achievement, which may seem impossible to attain without the use of evidence previously acquired through measurements. Hence, the risk of circularity (**Figure 1**), since answers to the questions "What counts as a measurement of X?" and "What is X?" often seem to presuppose one another when a theoretical understanding of the quantity or characteristic of interest is weak.[7] This means that the risk is more likely to occur when knowledge of the empirical relationship among the representing quantity and the represented quantity is yet in the making (van Fraassen, 2008).

Recent approaches in the epistemology of measurement have suggested that the circularity itself is not vicious, if we take a historical and coherentist approach (Chang, 2004; van Fraassen, 2008, cf. Tal, 2020). Rather than trying to avoid the risk of circularity, this should be embraced as a constitutive part of the process that leads to progress in measurement. According to these perspectives, the meanings of quantity concepts emerge from a historical and iterative process of mutual feedback between theoretical advances and improvements in measurement standards. With each iteration, the quantity concept is re-coordinated to

6 Chang (2004) labels this issue the "problem of nomic measurement" (cf. also Sherry, 2011; Bradburn et al., 2017), while van Fraassen calls it the "problem of coordination," following an epistemological tradition that dates back to the turn of the twentieth century (Mach, 1896/1986; Reichenbach, 1920. Cf. Padovani, 2017 for a discussion).

7 The picture provides a general description of an epistemic problem. This description abstracts away from the specific measurement system (i.e., the concrete measurement procedure and the theoretical model of the measurement process), as well as from the measurement target under investigation. For examples of epistemological analyses of this problem in different scientific disciplines, see the references in footnote 8.

a more stable set of standards, which allows for theoretical predictions to be tested more precisely. This, in turn, enables subsequent development of theory and the construction of more stable standards, and so on. Indeed, we can only realize how this process avoids vicious circularity when we look at it either "from above," i.e., in retrospect given our current scientific knowledge, or "from within," by looking at historical developments in their original context (van Fraassen, 2008: p. 122).

These recent coherentist approaches to measurement have developed from a primary focus on examples from physics, hand in hand with developments in metrological discussions also primarily targeting physical measurement and engineering (e.g., Mari, 2003; Frigerio et al., 2010; Giordani and Mari, 2012). One crucial feature of these approaches is that they shift from an exclusive focus on mathematical representational structures and the definition of quantity terms typical of classic mathematical theories of measurement, like the Representational Theory of Measurement. Instead, these approaches pay substantial attention to *realizations* (cf. Tal, 2020), that is, the physical instruments or procedures that approximately satisfy certain definitions of quantities (cf. JCGM, 2012: 5.1). These coherentist perspectives have been applied to analyze how measurement circularity can emerge and be tackled even beyond the physical sciences.[8]

Metrologists and psychometricians that are in dialog with these coherentist approaches have attempted to bridge physical and psychological measurement under overarching models of measurement (e.g., Mari et al., 2016, 2023). However, the very concept of a realization as provided by the JCGM, when translated into the context of psychological measurement, implies specific and difficult challenges that have received limited consideration by the philosophical and metrological literatures just mentioned. Two of these challenges are particularly relevant to the problem of circular measurement. The first concerns the fact that identifying empirical regularities which describe the relationship between two quantities or properties in a specific measurement interaction constitutes an intrinsic challenge for psychology.[9] The possibility to represent a characteristic that is not directly observable in terms of another observable property or quantity requires, in fact, an unbroken chain of interactions that goes from the first observable property to the measurand (JCGM, 2012). This chain of interactions is established through the identification of causal quantitative relations from the first property to the measurand. Most natural sciences can rely on shared perception as a criterion for metrological traceability, i.e., on the fact that inter-subjective agreement on what is being observed can be achieved, thus grounding the possibility to further infer causal empirical relationships among quantities. As the problem of measurement circularity shows, identifying these empirical regularities, also known as *measurement laws*, can be difficult in all sciences. While, as I will discuss, Fechner developed his quantification of sensation by adopting a standard in a spatio-temporal sense, this does not seem a viable possibility for a great part of psychology. This is mainly because its intraquestive measurement methods based on subject reports cannot support shared perception as a criterion for metrological traceability (Uher, 2019, 2020).

The second challenge concerns the fact that, in the psychological literature, realizations are often taken to refer to the questionnaires or other standardized assessment tools through which psychological measurement is performed. Therefore, according to this interpretation, it is the representational relationships among these measurement instruments, the target characteristics that they are supposed to be informative about, and the constructs that provide definitions of those characteristics, that are relevant to successful measurement. Indeed, this understanding has been for a long time at the center of discussions concerning validity, a key methodological notion for evaluating the quality of measurement and assessment tools in psychometrics.[10] The aspect of validity that, from the 1950s, started to be called *construct validity* involves building and testing theories about psychological characteristics which we also try to empirically access via measurement (Cronbach and Meehl, 1955; Messick, 1989).[11] One of the aims of construct validation is to clarify the definition of characteristics that are also measurement targets, so that the outcome of a certain measurement procedure can justifiably be claimed to be informative about the intended measurand, rather than about something else. Indeed, approaches based on construct validity resonate, to some extent, with the coherentist perspectives on measurement previously discussed, based as they are on a process of mutual refinement between measurement standards and theoretical concepts over time.

Yet, as both philosophers and methodologists have pointed out, conceptualizations of the relationship between theoretical constructs, the psychological phenomena that they describe, and the measurement outcomes that are supposed to be informative about them, remain underdeveloped in construct validity theories, thus leaving room for different interpretations of the meaning of test results.[12] In addition, the tendency to

8 These include, among others, medical measurement (McClimans, 2013), physical anthropology (Luchetti, 2022), perception studies (Barwich and Chang, 2015), and psychometrics (McClimans et al., 2017).

9 The reason for this difficulty is that the historical development of successful measurement procedures for a certain quantity or property is often intertwined with the empirical process of identification, confirmation and refinement of the relevant measurement laws that are required to infer information on the measurand from the result of a measurement process (Chang, 2004; Riordan, 2015; Luchetti, 2020). Yet, during calibration, i.e., the modeling of a measurement process, these empirical regularities are usually taken as fixed background presuppositions that justify the measurement inference. Therefore, the calibration and standardization of measurement procedures are often performed with only a partial knowledge of the necessary theoretical background (Barwich and Chang, 2015; Tal, 2017).

10 Validity as a technical term in this sense was first explicitly introduced in the context of attempts at standardizing intelligence testing in the 1920s, but it was progressively adopted as a methodological notion in domains beyond psychology and education. Even though validity in its original sense is commonly agreed to indicate the extent to which the assessment of an item is informative about the characteristic of interest, these developments led to a proliferation of validity concepts and taxonomies (cf. Newton and Shaw, 2014; Slaney, 2017). See, for example, Borsboom et al. (2004) and Markus and Borsboom (2013) for an overview of contemporary debates surrounding validity in psychometrics.

11 As of today, the unitary understanding of validity adopted, for instance, by US Standards in psychology and education is inspired by the construct validity perspective, even though it includes evidence from sources that were previously related to other validity notions (cf. American Educational Research Association et al., 2014).

12 See, for instance, Borsboom et al. (2009) for a criticism of construct validity from within psychometrics; Slaney and Garcia (2015) for a discussion of the use of "construct" language in psychology; Alexandrova and Haybron (2016) for a philosophical critique of the notion of construct validity; Stone (2019), Feest (2020), and Zhao (2023) for recent philosophical perspectives.

focus on questionnaires and standardized assessments as the only measurement instruments can lead to underappreciate the complex epistemic role of test subjects in the measurement interaction. Indeed, psychological measurement presents us with the peculiar issue of conceptualizing humans as both objects of measurement and measurement tools, thus challenging any approach to measurement which tries to dispense from a subjective evaluative component. Fechner was a forerunner of this realization, in a trajectory that—passing through Stevens' (1956) method of magnitude estimation based on the conception of the person as a measuring system—arrives at recent systematic perspectives on the "human as a measurement instrument" (e.g., Berglund et al., 2012; Pendrill and Petersson, 2016; Pendrill, 2019).[13]

A focus on the subjective component of measurement will be central to my analysis of Fechner's quantification of sensation and his approach to measurement circularity. Indeed, the recent coherentist epistemologies of measurement have reminded us that a human component is present in all measurement. This is because, at some point in all histories of quantification, inter-subjective evaluation, rather than reliance on well-established quantitative relations, was the basis for accepting certain measurement standards as valid. Therefore, such a consideration is most relevant in cases where the issue of measurement circularity is a challenge to the coherence of the assumptions on which quantification is based. By relying on a coherentist perspective of measurement, I will emphasize the "human" component of Fechner's approach to the quantification of sensation, which required him to put the subjective at the center of his quantification both methodologically and epistemologically.

## 3 Fechner's psychophysics and the making of sensation as a quantity

Initially trained as a medical doctor, Fechner [1801–1887] became a central figure in nineteenth-century German science and culture, contributing to several fields from physics to psychology, from statistics to esthetics, from metaphysics and the theory of mind to satirical literature (Fancher, 1996; Arendt, 1999; Heidelberger, 2004). Some narratives (e.g., Boring, 1961), characterize Fechner's psychophysics as an attempt to scientifically substantiate his philosophical view of the relationship between mind and matter, according to which the physical and the mental are two manifestations of one and the same reality (cf. Fechner, 1851/1957). Instead, several historians have emphasized the coherence of Fechner's psychophysical research program with his broader view of scientific inquiry (e.g., Marshall, 1982; Heidelberger, 2004). In addition, they have connected Fechner's emerging interest in psychophysics with central biographical events, such as his experience of prolonged visual deficiency and temporary mental impairment (e.g., Nicolas, 2002; Meischner-Metge, 2010).

Experiments on sensory modality had been performed from the seventeenth century, and psychophysical methods were systematically used in the work on touch carried out by Ernst Heinrich Weber [1795–1878]. Weber (1834, 1846) used comparisons between stimuli to identify thresholds of experience, that is, to identify the minimum stimulus required to perceive a sensation.[14] Among his results, Weber showed that the stronger a stimulus, the more intense should another stimulus be so that the difference with the former can be sensed. In other words, the minimal change in stimulus required for a difference in sensation to be perceived is a constant fraction of the values of the stimulus in the background. Therefore, the smallest discernable distinction between two stimuli can be expressed as an invariable ratio between them, independently of their strength. The formula expressing this ratio is: $\Delta R/R = c$, where $\Delta R$ is the relative threshold for the stimulus, that is, the limit at which the difference is discernible, R is the stimulus and $c$ a constant specific to each sensory modality.

Fechner invented the term *psychophysics* to refer to the scientific study of the functional relationship between body and mind, which he had intended to pursue as an exact science well before getting acquainted with Weber's empirical results (Marshall, 1982). Fechner conceived psychophysical processes as those physiological bodily processes immediately accompanying psychical events. Central to his psychophysical theory was the distinction between inner and outer psychophysics. Inner psychophysics focuses on the relation of the mental to the internal functions with which psychical activity is closely related, that is, on the relationship between the mental and neurophysiological activity. *Psychophysical excitation* was Fechner's term to describe the process, occurring in the brain and in the rest of the nervous system, through which the crossing of nerve tracts generated psychical activity. Outer psychophysics, instead, focuses on the relation of the mental to the body's external aspects, i.e., to the physiology of the senses.

Initially, Fechner searched for knowledge of the nervous system that would allow him to pursue inner psychophysics and, thus, directly investigate the causal processes giving rise to experience. However, he could not find such knowledge. The biophysicists working on the physical-chemical explanation of biological processes at the time were scarcely interested in the brain and the nervous system, plausibly because they did not view consciousness and higher mental activity as explainable in materialistic terms (Culotta, 1974). Therefore, Fechner's only viable empirical access to psychophysical processes was through the use of the indirect measurement methods offered by Weber's outer psychophysics, that is, the study of the relationship between physical stimuli and sensations. In this sense, Fechner conceived of psychophysical processes as abstract theoretical constructs when he wrote that "The mental intensity of an element is a mathematical fiction which has no other meaning than to provide for a calculation of a relationship which occurs in a system of elements" (Fechner, 1851/1957: p. 374). Yet, for this mathematical fiction to have

13 These perspectives aim to account for the fact that "screening and testing of participants as measuring instruments are absolutely necessary for reliable and valid psychological measurement" (Berglund et al., 2013), thus emphasizing an underappreciated dimension of analysis in the epistemology of psychological measurement. While the psychometric approach to measurement has developed fruitful tools to address this dimension, such as the Rasch model, this has been unevenly recognized within both psychology and the epistemology of measurement (e.g., McClimans et al., 2017).

14 The concept of threshold in psychology was introduced by Herbart (1824–1825), who defined it mathematically. For a historical overview of the notion of threshold, see Corso (1963).

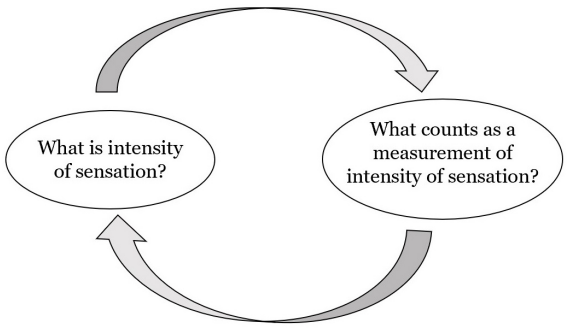



FIGURE 2
The epistemic circularity faced by Fechner in his attempt to quantify the intensity of sensation. In this case, the question "What is intensity of sensation?" includes a number of other sub-questions, including "What kind of property is intensity of sensation?" and "Is it quantifiable?"

concrete meaning Fechner had to establish a mapping between the characteristic of interest, viz., the intensity of sensation, and some measurable proxy. This mapping would ensure that his measurement methods would actually measure what he intended to measure in the absence of independent standards. Put it in another way, Fechner had to deal with the problem of epistemic circularity in measurement (**Figure 2**): How could he identify the "right" way of measuring intensity of sensation without already presupposing some quantitative understanding of intensity of sensation?

As I have mentioned, Weber had already established that some form of reliable measurement could be achieved in the experimental study of sensory thresholds for the different sensory modalities, by relying on the linear function relating physical stimuli and sensory thresholds that he identified. Indeed, his approach rested on identifying relative thresholds of experience based on increments of the stimulus, that is, on *ordering* sensations of different intensities according to the intensity of the stimulus that produced them. Fechner's goal was more ambitious, in that it aimed at *quantifying* sensations, based on his firm conviction that psychical phenomena have a quantitative dimension (Fechner, 1858). To this purpose, he set out to construct a mapping between the intensity of sensory stimuli, his only available physical proxy, and his attribute of interest, viz., experience, operationalized as intensity of sensation. This mapping required establishing (i) a measurement unit that could ground a scale of intensity of sensation, (ii) a functional relationship that would justify the representation of intensity of sensation in terms of intensity of the stimulus, and (iii) a material measurement standard that would embody this functional relationship and, thus, enable the actual quantitative study of experience.

Indeed, the only material measurement standard for which a functional relationship between stimulus intensity and sensation intensity could be identified, and that could then be used to measure sensation, is the human body. The very possibility of psychophysics as a quantitative discipline in Fechner's sense was based on the assumption of a shared human physiology, which ensures the stability of the functional dependence of sensory reaction from stimulus intensity. As I will show more in detail in section "4.2 Stabilizing the problem of measurement circularity," this assumption was a central component of Fechner's approach to measurement circularity.

Fechner's first important conceptual innovation concerns how he developed a unit of measurement by using the fact, experimentally established by Weber, that the smallest discernable distinction between two stimuli can be expressed as an invariable ratio between those stimuli ($\Delta R/R = c$).[15] More precisely, Fechner used this regularity to define a *just noticeable difference* (jnd), that is, the smallest difference in sensation that corresponds to the smallest perceptible change in stimulus. In such a way, the change in stimulus used to produce a difference in sensation can be taken as a standard, i.e., a physical proxy, to measure equal units of sensation intensity. In other words, this provides a definition of the unit of a scale of intensity of sensation, which Fechner calls the *Fundamentalformel*, or basic formula: $\Delta E = \Delta R/R\, c$, where $\Delta E$ is a just noticeable difference in sensation, while the equation expresses which intensity of stimulation corresponds to a unit of sensation. To construct a measurement scale out of this definition of a psychological unit, Fechner had to make two assumptions. The first is that all jnds are of equal magnitude, that is, that they produce the same change in sensation, independently of the base value of the stimulus. The second is that the jnds can be summated in the same way as material units. Both assumptions were later to be subject to strong criticism.

Yet, the basic formula is not by itself sufficient to ground a measurement scale of sensation intensity. To that purpose, Fechner needed to identify a functional relationship that, by specifying the number of jnds that make up *all* differences in sensation, would justify the representation of intensity of sensation in terms of intensity of the stimulus. To precisely characterize this relationship, which he later called the *Maßformel*, or measurement formula (also known as "Fechner's law"), and deploy it as a constructive principle for his scale of sensation, Fechner had to tackle the circularity problem. In other words, he had to somewhat justify that this measurement scale based on his chosen unit was actually measuring what it was supposed to measure. In the absence of independent support for his definition of the unit of sensation intensity, he set out to construct his measurement scale through a sort of bootstrapping process (Heidelberger, 2004). Having his basic formula, i.e., his definition of a unit of sensation intensity, in the background, Fechner first tested empirically the equality of sensation intensities through the method of adjustments, an experimental technique in which the test subject can adjust the intensity of the stimulus until it reaches a threshold and a just noticeable difference is perceived. Then, he statistically reduced individual aberrations in the evaluation of equality of differences (i.e., in the identification of thresholds).

The next step was to test sensations of different strength to identify which increase in stimulus is required to obtain an increase in sensation that is subjectively experienced to be identical to the others. The datapoints obtained in this phase were meant to enable Fechner to empirically validate his scale. To obtain

15 Fechner used this empirical result to construct a unit for a measurement scale of sensation intensity already in 1851, without referencing Weber. Only later he referred to this regularity as "Weber's law."

these datapoints, Fechner adopted the method of right and wrong cases, based on comparing the weight of two containers and discriminating between the two respective physical stimuli.[16] He only used himself as an experimental subject, but he corrected for the possibility of differences in subjective evaluation of the stimuli intensities. He did so by repeating the same comparisons several times, and then using a normal distribution to represent the probability of discriminating the stimuli. On top of being a great innovation at the time, this methodological point will be relevant to my discussion of Fechner's approach to measurement circularity in Section "4.2 Stabilizing the problem of measurement circularity."

Finally, Fechner expressed these datapoints as a monotone function between the increment of sensation found to be constant and the increment of stimulus required for it. In other words, Fechner moved from differences in sensation to *differentials*, i.e., infinitesimally small units of sensation. This move was necessary to express his measurement formula in logarithmic terms and use it to justify the measurement scale he constructed.[17] The resulting measurement formula, $E = z \cdot \log R$, expresses the functional relationship between values of the representing quantity, intensity of stimulus, and the represented quantity, intensity of sensation, thus justifying the use of intensity of stimulus as a proxy for measuring intensity of sensation.

# 4 Epistemological insights from Fechner's quantification of sensation

## 4.1 Objections to Fechner's quantification and developments after Fechner

Fechner's critics found several assumptions underlying his proposed quantification of sensation intensity to be highly problematic.[18] Most critics rejected the significance of the measurement formula for inner psychophysics and focused on its role for outer psychophysics. This was not taken lightly by Fechner, who wanted his measurement formula to be regarded as an empirical law of inner psychophysics (Marshall, 1982). According to Fechner, in fact, the measurement formula has a double character. On the one hand, his functional relationship between the intensity of the stimulus and the actual target of measurement, i.e., the intensity of sensation, is based on a unit of measurement that, even though resting on Weber's empirical regularity, *stipulates* the standard for measuring sensation. On the other, the measurement formula expressed, according to Fechner, the relation between psychophysical excitation, i.e., the physiological phenomenon causing sensation, and intensity of sensation. This part of the law remained theoretical, given that psychophysical excitation could not be empirically accessed. In addition, the assumption that led Fechner from Weber's law to his basic formula, that is, that all jnds can be considered as equal, was particularly contested already by Fechner's contemporaries, together with the assumption that units of sensation can be added to one another just in the same way as physical units (e.g., Tannery, 1875a,b; von Kries, 1882; James, 1890). These two criticisms, which came to be discussed together by the label of "quantity objection" (cf. Boring, 1950; Michell, 1999, 2012), emphasized the lack of independent empirical justification for the two assumptions just mentioned.[19]

Fechner's derivation of his measurement law and his empirical method of constructing a scale by concatenating experimentally estimated units (the jnds) eventually produced a schism between physicists and psychologists in the 1930s. Their divergent assessment of whether it is possible to make a quantitative estimate of sensory events in the absence of independent measures of sensation intensity eventually led to separate paths in the development and assessment of conceptualizations of measurement throughout the twentieth century (Berglund et al., 2013). While this separation was something that Fechner himself had attempted to break, developments within twentieth-century psychophysics showed the empirical limitations of Fechner's measurement standard. Crucially, Stevens (1956, 1957) established that Fechner's units of sensation, the jnds, cannot be considered to be uniformly equal, as Fechner postulated. Stevens adopted the method of fractionation, a method by which the subject judges whether one weight is half that of another, or one sound twice as loud as another, etc. By making comparisons between incremental assessments of jnds and sensory experiences through fractionation, he showed that the jnds are, in fact, not uniformly equal. Fechner's logarithmic formula was eventually replaced by Stevens' power law, resulting from a modification of the basic formula (Stevens, 1969, 1970).[20] While the compatibility of Fechner's logarithmic formula with Stevens' power law and further formulations has been, and still is, a topic of debate in psychophysics (e.g., Wasserman et al., 1979; Laming, 1991, 2010), it became clear that Fechner's measurement formula is only applicable to a restricted range of sensory modalities. Even though Fechner's methods have never really been abandoned (e.g., Luce and Edwards, 1958; Eisler, 1963; Falmagne, 1971; Murray, 1993), later developments downsized the validity of Fechner's measurement standard and questioned the view of psychophysics as an enterprise aimed at discovering fundamental quantities (cf. Luce, 1972).

In the rest of this section, I will provide an assessment of Fechner's approach to measurement circularity by situating it in a historical perspective and in relation to his conceptualization of measurement. This will enable me to provide an assessment of his psychophysical project by looking at it both "from above," i.e.,

16 This method became later known as the method of constant stimulus (cf. Brown and Thomson, 1921; Guilford, 1936).

17 For a discussion of the epistemological implications of this modeling assumption, see Briggs (2021: p. 39–41).

18 For a detailed account of the criticisms against Fechner's quantification of sensation from his contemporaries, see Heidelberger (2004: p. 207–234). Cohen's neo-Kantian objection to Fechner's quantification and its impact on subsequent neo-Kantian philosophy are discussed by Giovanelli (2017). Feest (2021) reviews the objections raised by Gestalt psychologists against Fechner's additivity assumption. Biagioli (2023) discusses the relationship between Fechner's quantification and Helmholtz's view of sensory measurement.

19 See Briggs (2021: p. 51–55) for an excellent discussion of the quantity objection and its implications.

20 The first identification of a power relationship for the dependence of visual acuity on the intensity of the light by which the stimulus pattern was illuminated dates back to the work of Tobias Mayer in 1754 (Grüsser, 1993).

in retrospect given our current knowledge, and "from within," by considering historical developments in their original context.

## 4.2 Stabilizing the problem of measurement circularity

In section "3 Fechner's psychophysics and the making of sensation as a quantity," we have seen that Fechner's quantification of sensation intensity required presupposing a host of assumptions that were, at least at the time, untested or untestable. These included the assumptions concerning the equality and additivity of jnds, that became the focus of heated debates and are still relevant to methodological discussions today. However, much less attention has been paid to another of Fechner's assumptions, which had a crucial role both in his experimental practice and in his approach to the problem of measurement circularity. This is the assumption that all human individuals share a common physiology.

Fechner's approach to quantifying sensation involved using Weber's experimental methods of outer psychophysics, which relate behavioral response data to physical stimuli, in order to gain access to inner psychophysical processes, i.e., the neurophysiological goings-on of sensory experience. The possibility of this methodological jump was justified by Fechner's assumption of a shared human physiology. For the purposes of establishing the correlation between the mental and the physical, in fact, Fechner considered that the individual differences in the physiological make-up of test subjects were irrelevant. In addition, this assumption justified the possibility to use himself as one of few, or even the only, test subjects in his experimental practice. Epistemologically speaking, this idealizing assumption replaced the process of standardizing his measurement instrument, i.e., the human sensory apparatus.

More generally, the assumption of a shared human physiology ensured the stability of the empirical regularity black-boxed by his measurement formula, i.e., the causal relationship between the intensity of a sensory reaction and the psychophysical excitation produced by a stimulus of a certain intensity. Fechner was aware that subjective evaluation has an impact in the identification of thresholds of experience, in that it provides an important source of variability. For this reason, he characterized the notion of threshold in statistical terms.[21] As I previously mentioned, Fechner replicated the experiments through which he established the empirical datapoints validating his measurement formula. This methodological step allowed him to control for differences in subjective judgment of the stimuli. Yet, as he was using only himself as a test subject, this step could not control for possible differences in physiological make-up. The assumption of a shared human physiology *de facto* enabled Fechner to discount the possibility of experimental variation resulting from differences in psychophysical excitation due to different neurophysiological make-ups of test subjects, the impact of which, as we have seen, would anyway be out of reach given the state of neurophysiological knowledge at his time. By anchoring the reaction to sensory stimuli to a univocal and stable causal basis, i.e., our shared sensory apparatus, Fechner could then set out to develop a representational mapping between the empirically accessible side of the functional relationship that he aimed to establish, i.e., the intensity of the stimulus, and the characteristic that was his actual measurement target, i.e., the intensity of sensation.

In this sense, the assumption of a common human physiology has a special epistemic status, since it provided Fechner with an anchor to keep the circularity problem stable. Without this assumption, the variability due to individual differences in physiological make-up would have made it much more difficult, if not impossible, to establish the functional dependence between intensity of stimulus and intensity of sensation. This is because, if that were the case, differences in reactions to the same sensory stimulus would have been considered as partly dependent on physiological differences among subjects. Yet, there would have hardly been a way to factor the extent of the causal influence due to these differences, given the insufficient neurophysiological knowledge of the time.

## 4.3 Reassessing Fechner's standard in light of his epistemology of measurement

In addition to the assumption of a shared human physiology, the very idea that sensation itself is something that can be at all quantified was another crucial untested assumption behind Fechner's approach to measurement circularity. Fechner's conventional assumption of the equality and additivity of jnds has been directly invoked as the remote cause of the overly liberalized current view of quantification in psychometrics (Michell, 2006, 2008). From this perspective, Fechner stipulated his measurement standard without securing a logically prior step. That is, he did not verify empirically the quantitative character of the relationship between the characteristic of interest, i.e., the intensity of sensation, and the chosen standard, i.e., the intensity of stimulus (Michell, 2006, 2012). While engaging with this argument is beyond the scope of this contribution, in my view we can understand Fechner's assumption of the quantifiability of sensation only against the backdrop of his nuanced epistemological perspective and from within the historical context of his measurement practice.

Some commentators have emphasized how Fechner's approach to quantifying sensation was entangled with his correlative interpretation of measurement (Murray, 1993; Heidelberger, 2004; Briggs, 2021). According to Fechner, in fact, the relationship between the external stimulus and sensation is not a causal one. While the stimulus causes psychophysical excitation in the brain or in the nervous system, it is not directly causally related to sensation. Rather, the stimulus is only functionally linked to sensation, inasmuch as it is used as a representation of the latter.[22] The possibility to represent intensity of sensation in terms of the intensity of the stimulus is warranted by the mapping expressed by the measurement formula, which describes

21 This is a move that he had already made when conducting his inquiry on Ohm's law and the Galvanic circuit (Fechner, 1831; cf. also Marshall, 1990).

22 The conventionality of this move leads to define new units for the physical stimulus, a result that was criticized by Boring (1921).

the relationship between these two quantities with respect to a concrete measurement system, that is, the human body. The choice of intensity of stimulus as the other term of the functional-representational relationship is indeed a conventional one, but the choice of a convention is only a part of the story. Indeed, Fechner's measurement formula established a correlation between the intensity of stimulus taken as a representing quantity, and the intensity of sensation as the represented quantity. Yet, in Fechner's view, the importance of the measurement formula went beyond a mere correlational aspect. From the perspective of his inner psychophysics, as we have seen, the measurement law was itself justified by the causal relationship between psychophysical excitation and intensity of sensation, a relationship that was yet to be empirically discovered.

The innovative character of Fechner's correlative view of measurement had an influence that went well beyond the field of psychophysics. Notably, Fechner's correlative view was taken by the physicist Ernst Mach as a blueprint for his own view of measurement (Heidelberger, 1993, 2004, 2010; Briggs, 2021; Staley, 2021). In Mach's (1896/1986) view, measuring does not amount to discovering a state of the matter, but rather to discovering the relation holding between the measured characteristic and a chosen measurement standard.[23] Particularly in the early stages of developing measurement procedures, the choice of measurement instruments and standards is conventional and guided by pragmatic considerations. Yet, by putting some sort of measurement standard in place, it enables the collection of empirical data that then allow for further empirical investigation of the relationship among the quantities that was somewhat postulated in the first place. This relational and iterative understanding influenced, in more recent times, the coherentist perspectives on measurement progress that I introduced in section "2 Epistemic circularity and psychological measurement," especially Chang's (2004) view of progress through epistemic iteration.

Before turning to my assessment of Fechner's approach vis-à-vis subsequent developments in psychophysics, I must address two further points. The first is that the transition toward quantitative science that was characterizing German and, more generally, European science at the time constituted a central influence on Fechner's approach to measurement. Most importantly, Fechner was working within the so-called *Euclidean* tradition of measuring magnitudes, according to which "ratios of magnitudes are equal to ratios of natural numbers or are approximated by ratios of natural numbers" (Zudini, 2011: p. 76).[24] In other words, Fechner's underlying conception of measurement was shaped by this classical understanding of measurement, by which all measurement requires quantification on a ratio scale, thus necessitating an absolute zero point and equality of intervals among units of the measurement scale.[25] Therefore, the Euclidean model constrained the range of possible measurement scales that Fechner could choose to develop his measurement scale, and inevitably led him to strive for a quantitative approach that would enable him to measure intensity of sensation on a ratio scale.

Second, it is crucial to emphasize that Fechner's epistemic goal, much as it was shaped by the search for precise quantification, was not that of discovering the *ultimate* quantitative model of human sensory experience. In fact, Fechner used mathematical tools not only with the aim of representing quantitative relationships, but also as investigative tools, for example in the case of his statistical notion of threshold (Marshall, 1982). In this respect, mathematization was certainly a goal for Fechner, but not in the sense of providing a quantitative description of human experience that would not require further refinement. This is demonstrated by the fact that Fechner was very much aware of the provisional character of his quantification, since he regarded his *Elemente* more as a research progress report than as a final scientific product (Fechner, 1860, vii). In addition, in his treatise he recognizes the absence of practical alternatives to taking the intensity of the stimulus as a concrete standard to quantify sensation, and he emphasizes that the main role of theorizing is its function of generating testable assumptions, rather than of providing incontrovertible definitions. All these points suggest that his goal of achieving mathematical tractability for the supposed quantitative phenomenon under investigation was very much open to the possibility of refining his formal characterization through empirical considerations made available by further investigations. Indeed, to establish a standard for measuring sensation intensity Fechner had to resort to a number of conventional choices, most notably that of the equality of jnds. Yet, it was very clear to him that these specific choices were only pragmatically necessary and that they were revisable in the light of empirical evidence.

In short, the ideal of universal quantification spreading fast in the nineteenth century science pushed Fechner toward the goal of providing an overarching model of quantification of sensation. This required embracing core untested assumptions, such as the one concerning the quantifiability of sensation, that were modeled on physical quantification. Fechner developed an original approach to devise a measurement standard based on his correlational view of measurement. This approach required him to make untested assumptions about the quantitative structure of the characteristic of interest, in order to overcome the dead-end of circularity. The non-testability of the causal complement to his correlative measurement formula and the issues raised by the quantity objection are indeed crucial unresolved aspects of his approach to measurement. Yet, several features of Fechner's epistemic attitude, such as his recognition of the absence of practical alternatives to his chosen standard and of the revisability of his standards based on empirical considerations, show the modernity of his epistemological standpoint. Viewed from this perspective, Fechner's own approach to measurement seems to resonate with more recent approaches to construct validity in psychological measurement and coherentist views in epistemology of measurement. This

23 When Mach (1872/1909) urged that the proper aim of science is to discover the fixed functional dependence of phenomena on one another, he was following the lead of Fechner (cf. Ryckman, 1991).

24 For brief characterizations of the Euclidean tradition and its historical significance see, for instance, Mari (2013).

25 This view of measurement was to be abandoned by subsequent approaches to psychological measurement while a strict identification between measurement and quantification has largely been discarded as of today.

is because these approaches emphasize that progress along any of the interacting dimensions of theory, experimentation, and measurement should reverberate on the network of assumptions and empirical generalizations involved in the definition of quantities and units.

### 4.4 Fechner's standard as a first epistemic iteration for psychophysical measurement

In the previous paragraphs, we have seen that the empirical validity of Fechner's quantification of sensation has been rescaled in the light of twentieth century developments in psychophysics. Most importantly, while his assumption of the equality of the jnds was empirically disproved, his logarithmic measurement law was found to hold only for a restricted range of sensory modalities. Therefore, from our vantage point, Fechner's overall project of quantifying sensation might be regarded as an unsuccessful enterprise. Yet, if we take a view from "above," a different assessment is possible.

First of all, Fechner's methods of experimentation and statistical analysis, through which he located the jnds and assessed the sensitivity of human discrimination, were universally adopted (Stigler, 1986; Briggs, 2021). In addition, several commentators have emphasized that Fechner's construction of a measurement standard for intensity of sensation actually enabled the subsequent advancement of psychophysical measurement (e.g., Falmagne, 2002; Heidelberger, 2004; Isaac, 2013). Fechner's way out of the circularity issue made it possible to treat psychophysical data mathematically, thus enabling scientists to gather more empirical knowledge and develop more advanced measurement techniques, such as multidimensional scaling (cf. Isaac, 2013, 2017). This, in turn, enhanced the empirical investigation of the quantitative relationships among jnds and made it possible to replace Fechner's standards in light of empirical considerations. More precisely, the fact that Fechner put a measurement standard in place opened the door for the mathematical analysis of psychophysical data. This enabled the generation of precise predictions about just noticeable differences, which could then be empirically tested, thus enabling the refinement of the measurement standard itself at a later stage.

In sum, Fechner's engagement with the issue of measurement circularity led him to a quantification of sensation that achieved sufficient mathematical tractability to start off a long-term process of refinement of the measurement standards for intensity of sensation over time. The measurement outcomes obtained through Fechner's quantification were, in fact, taken as the empirical basis of a process that, in the *longue durée*, enabled the study of the quantitative relationships among jnds and led to the development of more accurate standards, thus making psychophysics the empirically successful research program that is today. From this point of view, Fechner's quantification can be considered as successful insofar as it satisfied the goal of providing a first measurement standard for sensation intensity, even if its empirical adequacy was later found to be limited. In this respect, Fechner's standard represents a first epistemic iteration in the process of developing psychophysical measurement. His approach to the problem of measurement circularity, with its strengths and limitations, served the purpose of overcoming an impasse and providing a first, temporary standard which could then be refined over time.

## 5 The relevance of Fechner to current methodology of psychological measurement

So far, I have analyzed the approach to the epistemic circularity behind Fechner's quantification of sensation. I have discussed how Fechner stabilized the circularity by making a number of assumptions, which concerned the subject matter that he was attempting to quantitatively model, its relationship with a spatio-temporally located standard, and the notion of measurement itself. I contextualized the development of Fechner's quantification from within the framework of nineteenth-century science, and I emphasized that the consolidating ideals of quantitative objectivity and universality were built into his creation of a measurement standard for sensation intensity. Nevertheless, I stressed that his approach to measurement, and to the circularity issue specifically, had innovative aspects, which resulted from Fechner's appreciation of the subjective aspect of measurement, both methodologically and epistemologically.

Finally, I have shown Fechner's contribution can fit a story of success, to the extent that we regard his approach to circular measurement as conducive to a first, albeit imperfect, standard for measuring sensation, which could start a process of epistemic iteration. Taking this perspective seems also justified by the relationship of Fechner's own conceptualization of measurement with coeval perspectives that were embracing some form of coherentism about measurement. In addition to Fechner's influence on Mach, Briggs (2021) emphasizes that Fechner was working at a time in which Maxwell and Thomson (also known as Lord Kelvin) were actively reflecting on how advancements in physical measurement are carried forward by the identification of the proper measurement laws. In this sense, Maxwell and Thomson were envisioning a coherent system of fundamental and derived units defined by referring to a set of constants of nature, thus preconizing the approach currently taken by the International System of Units (cf. de Courtenay, 2022).

In the context of psychology, it has been argued that this "Maxwellian" approach has been insufficiently considered by methodologists of measurement, to the benefit of traditions such as operationalism and representationalism (McGrane, 2015). Fechner was himself a forerunner of this approach, in that he developed his measurement standard by identifying a measurement formula that functionally related internal sensation to a spatio-temporal property, i.e., the intensity of the physical stimulus. Indeed, his formula was only correlational, since the functional relationship was not based on an empirical causal law, but only on a statistically modeled set of observations used to infer the magnitude of the characteristic of interest. As we have seen, however, the causal law was, in his view, to be eventually identified empirically by research in inner

psychophysics, which would provide the final validation to the relationship underlying his measurement standard. While this validation has not been provided, this aspect of Fechner's approach seems to be the carrier of an optimistic vision of measurement, "one that reflects the ongoing efforts to uncover and understand the causal mechanism underlying the relationship" (Briggs, 2021: p. 52).

Even if we grant that Fechner's vision may hold for psychophysics, the approach based on identifying an empirical causal law that justifies the representational relationship in measurement might not be regarded with the same optimism in most of quantitative psychology. This is because, "[c]ontrary to beliefs widespread in psychology, findings about individuals' perceptions of physical phenomena cannot be generalized to all psychical phenomena, which, given their non-spatial properties, differ fundamentally from the spatially extended phenomena the perception of which is studied in psychophysics" (Uher, 2019: p. 242). In other words, the fact that there are no evident observable properties that can be linked to the psychological characteristics that we aim to measure may be considered as an intrinsic barrier to this approach. This is because most psychological instruments are not based on the detection of some perceptible quality, as in the case of sensations of physical stimuli. Instead, they are necessarily based on language, thus involving interactions between the human instrument (i.e., the rater), the non-human instrument, and the phenomena and properties under investigation. Interpretive decisions, rather than empirical causal relationship, are therefore required to establish a representational relationship in these measurement systems (Uher, 2021a).

The question then is the following: To what extent, if at all, can we apply insights from Fechner's psychophysical measurement to current quantitative psychology? On the one hand, his approach to the development of a measurement standard has been praised for its radically innovative epistemological import. On the other, it has been regarded as intrinsically flawed or unsuitable to most needs of quantitative psychology. In my view, the relevance of Fechner for current issues in quantitative psychology should be searched neither in his specific way of developing a measurement standard for sensation, nor in his theory of measurement *per se*. As such, we cannot take the success story of psychophysics, and of Fechner's role in it, as grounds for optimism with respect to the possibility of achieving a similar form of quantification in the rest of psychology. Yet, the strengths and limitations of his general epistemic attitude toward measurement can provide important reflections for current quantitative psychology. Most importantly, his approach toward the problem of measurement circularity gives us the possibility to rethink important epistemic categories central to the assessment of measurement in current psychology, such as the notion of successful measurement and the notions of validity and objectivity.

A first point concerns the goal of stabilizing the circularity. As I have shown, in Fechner's approach, the assumption of a shared human physiology functioned as an essential anchor to achieve stabilization. The presence of an idealizing element opens up a question concerning both the justification for this idealization and its implications. Clearly, this assumption was not taken for granted in other contexts of psychophysical research at later stages, whereby *differences* among sensory experiences of individuals, rather than their similarities, became relevant. For instance, this occurred when it started becoming clear that "individual variations in sense experience approached but did not quite align with the new biological theories of human variation powered by the concept of heredity" (Fretwell, 2020: p. 3). In this sense, while idealizing assumptions such as this one might be necessary to stabilize the circularity, it is crucial to clearly identify the scope of their justification. This is very relevant to psychological measurement in general, inasmuch as a certain measurement tool is aimed, for instance, at tracing differences within populations (e.g., distinguishing among human groups according to personality traits). When the goal is that of identifying differences, rather than broad generalizations, it becomes very difficult to find justification for such strong idealizations.

Most importantly, Fechner's need to stabilize the circularity derived from his epistemic goal of identifying a first measurement standard for sensation intensity. This, in turn, involved a trade-off of epistemic values, which is relevant to the assessment of what counts as successful measurement. The adoption of idealizing assumptions about the measurand, the measurement instrument, and their relationship, is always required to model the measurement process (Tal, 2017). These idealizations serve the purposes of model tractability, but this occurs to the detriment of the possibility of achieving complete representational accuracy, which is itself an idealization (Teller, 2013, 2018). By assuming a shared human physiology, Fechner could dispense with accounting for individual neurophysiological variability of test subjects and could experiment mostly on himself, thus privileging generality of representational accuracy.

The insight that can be taken from Fechner's use of untested or untestable idealizations is that these assumptions can be necessary to stabilize the circularity and enable the development of a new measurement standard. As such, these assumptions can be crucial to successful measurement, where success should be understood as relative to the purpose at hand and to the trade-off of epistemic values that it underlies. In the case of Fechner, the use of this idealization was conducive to achieving mathematical tractability of psychophysical phenomena. This achievement was not itself sufficient with respect to the overarching goal of providing a universal, empirically adequate quantitative representation of intensity of sensation, but it did enable the improvement of the measurement standard at a later stage of psychophysics. Yet, Fechner did not explicitly acknowledge the use of this idealized assumption with reference to the achievement of a specific goal, nor did he clearly acknowledge the validity of the resulting measurements as context-dependent.

This insight can be relevant to current quantitative psychology quite independently from the stance concerning where research efforts should be directed to improve the standards of psychological measurement. Indeed, many voices have pleaded for more and better theorizing in psychology in the wake of the replication crisis (e.g., Muthukrishna and Henrich, 2019; Oberauer and Lewandowsky, 2019). However, how should this plea be tailored to address measurement circularity in psychometrics? Among the challenges of psychometrics, we find, for example, the fact that standardized questionnaire statements are interpreted very differently across test subject, and even by the same subject in different circumstances (e.g., Lundmann and Villadsen, 2016).

If we take the empirical identification of a causal quantitative relationships underlying the existing measurement standards as our goal (e.g., Kellen et al., 2021), then we should strive for a better theoretical and causal understanding of the measurement instruments used, in particular the language-based reports with which psychology cannot dispense (Uher, 2021b). Indeed, the multi-dimensionality and instability of the psychological subject matter, as well as the availability of intraquestive methods only, call for searching something quite different from the single causal law that Fechner thought could justify his measurement standard. For example, an important contribution in this direction would be to better identify which conditions affect the interpretations given by test subjects (i.e., the humans as measurement instruments) to the items of standardized assessment tools, and how this feeds back into converting resulting information into fixed scales.[26] While such an effort is made, however, current standards from which such research is conducted would still presuppose idealized untested or untestable assumptions about the causal quantitative relationships. The story of psychophysics and Fechner tells us that these assumptions can play an important role in the long run, but that their scope of application and impact on the validity of measurement must be carefully assessed, especially by making explicit the measurement goals and the related value trade-offs.

The second, related point concerns the categories of validity and objectivity that result from such a picture. Indeed, coherentist epistemologists of measurement have suggested that "quantitative structure is ultimately established through a coherentist fit between substantive theory and data that leads to improvements in various desiderata such as the scope, accuracy, and fruitfulness of the relevant inquiry. The process of establishing such coherence involves bottom-up discovery of relations in data alongside top-down, theory-driven corrections to the data" (Tal, 2021: p. 735). In other words, the process of refinement of measurement standards over time involves the progressive establishment of quantitative structure through coordinated improvements at the level of theory and of data which, in turn, can be evaluated as improvements thanks to reference to certain values. As I mentioned, the identification of quantitative structure would occur differently in psychometrics compared not only to the natural sciences, but also to psychophysics. In this sense, coherentism can be a helpful epistemological approach for quantitative psychology, if the search of a coherentist fit is not merely mimicked from paradigmatic cases in the physical sciences, but it is adequately paraphrased to the context of psychological measurement.

Most importantly, a proper characterization and focus on measurement circularity should be central to efforts in this direction. To understand exactly how, the story of Fechner's measurement standard for subsequent psychophysics reminds us of two important points. First, that quantification is open-ended, since it will always be possible to perfect the knowledge of quantities and of the relationships among them as science further progresses (cf. Riordan, 2015). Second, that the epistemic goals of quantification change over time, in parallel with changes and improvements in measurement standards and techniques, thus changing, in turn, the criteria for evaluating what counts as a successful, adequate, or useful form of measurement or quantification.[27] Therefore, rather than considering the circularity as an issue to be solved once and for all by means of an ultimate standard, methodologists should "listen" to what methodological and epistemological questions emerge in connection with the appearance of a specific form circularity in a specific context of inquiry. This, in turn, would open up questions concerning the values (epistemic or not) that are embedded in a certain measurement practice and the related trade-offs which, as we have seen, are related to the goals that are pursued by trying to achieve a certain, temporary solution to the measurement circularity.

In this sense, focusing on measurement circularity can be conceived as a hermeneutic tool (McClimans, 2023), which does not serve the only purpose of identifying rigid causal relationships that justify the quantitative structure of a measurable characteristic, at least in the short run. Rather, this tool is useful to reflect on the conditions of scope, accuracy and fruitfulness of a certain measurement standard, in other words, on the criteria of success that a scientific community wants to pursue by finding a certain, temporary solution to the circularity. In this sense, acknowledging the trade-offs of values can also mitigate some limitations of coherentist approaches when applied to subjective evaluations and the human sciences (cf. Thompson, 2023). By explicitly identifying what a certain solution to measurement circularity does and does not fulfill, the purposivity and selectivity of a measurement standard is acknowledged and the relative validity of the resulting measurements fully recognized. Such an understanding of validity as context-relative and purpose-oriented is very much in line with current standards (American Educational Research Association et al., 2014), and indeed it calls for a notion of objectivity that, when applied to measurement, will look quite different from the one that was guiding Fechner. By putting the subject back in the measurement process, Fechner initiated a process that, despite his convictions, requires us to acknowledge and integrate goals and values to objectively evaluate our measurement of the human.

## Author contributions

ML: Writing – original draft, Writing – review & editing.

## Funding

The author declares that financial support was received for the research, authorship, and/or publication of this article.

26 A recent example in this sense can be found in recent research on neurodegeneration, which has formulated metrological references for cognitive task difficulty that can be used to calibrate the measurement system function (Melin et al., 2021).

27 For an example of a recent perspective that emphasizes the relevance of goal-orientedness of quantification, and measurement in general, see Irribarra (2021, especially Ch. 4).

Open Access funding for this article has been provided by the Max Planck Institute for the History of Science.

## Acknowledgments

I would like to thank colleagues and affiliated members of the Max Planck Research Group "Practices of Validation in the Biomedical Sciences" for their continuous support and feedback. In particular, I thank Simon Brausch, Sam Ducourant, Alfie Freeborn, Ariane Hanemaayer, Lara Keuck, and Hanna Worliczek. I also thank Alistair Isaac for very helpful comments on earlier drafts of this manuscript.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Alexandrova, A., and Haybron, D. M. (2016). Is construct validation valid? *Philos. Sci.* 83, 1098–1109. doi: 10.1086/687941

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Arendt, H.-J. (1999). *Gustav Theodor Fechner: Ein deutscher Naturwissenschaftler und Philosoph im 19. Jahrhundert*. Frankfurt am Main: Lang.

Barwich, A. S., and Chang, H. (2015). Sensory measurements: Coordination and standardization. *Biol. Theory* 10, 200–211. doi: 10.1007/s13752-015-0222-2

Berglund, B., Rossi, G. B., Townsend, J. T., and Pendrill, L. R. (2012). *Measurement with Persons: Theory, Methods, and Implementation Areas*. New York, NY: Psychology Press.

Berglund, B., Rossi, G. B., and Wallard, A. (2013). "Measurement across physical and behavioral sciences," in *Measurement with Persons*, eds B. Berglund, G. B. Rossi, J. T. Townsend, and L. R. Pendrill (New York, NY: Psychology Press), 1–25.

Biagioli, F. (2023). Hermann von Helmholtz and the quantification problem of psychophysics. *J. Gen. Philos. Sci.* 54, 39–54. doi: 10.1007/s10838-022-09605-6

Boring, E. (1921). The stimulus-error. *Am. J. Psychol.* 32, 449–471. doi: 10.2307/1413768

Boring, E. (1950). *A History of Experimental Psychology*, 2nd Edn. New York, NY: Appleton-Century-Crofts.

Boring, E. G. (1961). Fechner: Inadvertent founder of psychophysics. *Psychometrika* 26, 3–8. doi: 10.1007/BF02289680

Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., and Franić, S. (2009). *The End of Construct Validity: The Concept of Validity: Revisions, New Directions and Applications*. Charlotte, NC: IAP Information Age Publishing.

Borsboom, D., Mellenbergh, G. J., and Van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bradburn, N. M., Cartwright, N., and Fuller, J. (2017). "A theory of measurement," in *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, ed. L. McClimans (London: Rowman & Littlefield), 73–88.

Briggs, D. C. (2021). *Historical and Conceptual Foundations of Measurement in the Human Sciences: Credos and Controversies*. New York, NY: Routledge. doi: 10.1201/9780429275326

Brown, W., and Thomson, G. H. (1921). *The Essentials of Mental Measurement*, 2nd Edn. Cambridge, MA: Cambridge University Press. doi: 10.1037/11188-000

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press. doi: 10.1093/0195171276.001.0001

Corso, J. (1963). A theoretico-historical review of the threshold concept. *Psychol. Bull.* 60, 356–370. doi: 10.1037/h0040633

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52:281. doi: 10.1037/h0040957

Culotta, C. A. (1974). German biophysics, objective knowledge, and romanticism. *Hist. Stud. Phys. Sci.* 4, 3–38. doi: 10.2307/27757326

de Courtenay, N. (2022). On the philosophical significance of the reform of the international System of Units (SI): A double-adjustment account of scientific enquiry. *Perspect. Sci.* 30, 549–620. doi: 10.1162/posc_a_00397

Eisler, H. (1963). Magnitude scales, category scales, and Fechnerian integration. *Psychol. Rev.* 70:243. doi: 10.1037/h0043638

Falmagne, J. C. (1971). The generalized Fechner problem and discrimination. *J. Math. Psychol.* 8, 22–43. doi: 10.1016/0022-2496(71)90021-6

Falmagne, J. C. (2002). *Elements of Psychophysical Theory*. Oxford: Oxford University Press. doi: 10.1093/oso/9780195148329.001.0001

Fancher, R. E. (1996). *Pioneers of Psychology*. New York, NY: WW Norton & Co.

Fechner, G. T. (1831). *Massbestimmungen über die galvanische Kette*. Leipzig: F. A. Brockhaus.

Fechner, G. T. (1858). Das psychische Maß. *Z. Philos. Philos. Kritik* 32, 1–24.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf u. Härtel.

Fechner, Z. (1851/1957). "Kurze Darlegung eines neues Principes mathematischer Psychologie," in *Zend-Avesta: oder über die Dinge des Himmels und des Jenseits vom Standpunkt der Naturbetrachtung (Vol. 1)*, ed. E. Scheerer (Leipzig: Voss), 203–207.

Feest, U. (2017). Phenomena and objects of research in the cognitive and behavioral sciences. *Philos. Sci.* 84, 1165–1176. doi: 10.1086/694155

Feest, U. (2020). Construct validity in psychological tests–the case of implicit social cognition. *Eur. J. Philos. Sci.* 10:4. doi: 10.1007/s13194-019-0270-8

Feest, U. (2021). Gestalt psychology, frontloading phenomenology, and psychophysics. *Synthese* 198(Suppl. 9), 2153–2173. doi: 10.1007/s11229-019-02211-y

Feest, U. (2022a). "Progress in psychology," in *New Philosophical Perspectives on Scientific Progress*, ed. Y. Shan (London: Routledge), 184–203. doi: 10.4324/9781003165859-13

Feest, U. (2022b). Data quality, experimental artifacts, and the reactivity of the psychological subject matter. *Eur. J. Philos. Sci.* 12:13. doi: 10.1007/s13194-021-00443-9

Fretwell, E. (2020). *Sensory Experiments: Psychophysics, Race, and the Aesthetics of Feeling*. Durham: Duke University Press. doi: 10.1515/9781478012450

Frigerio, A., Giordani, A., and Mari, L. (2010). Outline of a general model of measurement. *Synthese* 175, 123–149. doi: 10.1007/s11229-009-9466-3

Giordani, A., and Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Trans. Instrum. Meas.* 61, 2144–2152. doi: 10.1109/TIM.2012.2193695

Giovanelli, M. (2017). The sensation and the stimulus: Psychophysics and the Prehistory of the Marburg School. *Perspect. Sci.* 25, 287–323. doi: 10.1162/POSC_a_00244

Grüsser, O. J. (1993). The discovery of the psychophysical power law by Tobias Mayer in 1754 and the psychophysical hyperbolic law by Ewald Hering in 1874. *Behav. Brain Sci.* 16, 142–144. doi: 10.1017/S0140525X00029332

Guilford, J. P. (1936). *Psychometric Methods*. New York, NY: McGraw-Hill.

Hanfstingl, B. (2019). Should we say goodbye to latent constructs to overcome replication crisis or should we take into account epistemological considerations? *Front. Psychol.* 10:1949. doi: 10.3389/fpsyg.2019.01949

Heidelberger, M. (1993). Fechner's impact for measurement theory. *Behav. Brain Sci.* 16, 146–148. doi: 10.1017/S0140525X00029368

Heidelberger, M. (2004). *Nature from within: Gustav Theodor Fechner and his Psychophysical Worldview*. Pittsburgh, PA: University of Pittsburgh Press. doi: 10.2307/jj.10984425

Heidelberger, M. (2010). Functional relations and causality in Fechner and Mach. *Philos. Psychol.* 23, 163–172. doi: 10.1080/09515081003727400

Herbart, J. F. (1824–1825). *Psychologie als Wissenschaft, neu gegrundet auf Erfahrung, Metaphysik und Mathematik*. Königsberg: Unzer.

Hornstein, G. A. (1988). "Quantifying psychological phenomena: Debates, dilemmas, and implications," in *The Rise of Experimentation in Modern Psychology*, ed. G. Morawski (New Haven: Yale University Press), 1–34.

Irribarra, D. T. (2021). *A pragmatic Perspective of Measurement*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-74025-2

Isaac, A. (2017). Hubris to humility: Tonal volume and the fundamentality of psychophysical quantities. *Stud. Hist. Philos. Sci.* 99, 65–66. doi: 10.1016/j.shpsa.2017.06.003

Isaac, A. M. (2013). Quantifying the subjective: Psychophysics and the geometry of color. *Philos. Psychol.*26, 207–233. doi: 10.1080/09515089.2012.660139

James, H. (1890). *Principles of Psychology*. New York, NY: Holt. doi: 10.1037/10538-000

JCGM (2012). *International Vocabulary of Metrology—Basic and general concepts and associated terms (VIM)*. Paris: International Organization of Legal Metrology

Kellen, D., Davis-Stober, C., Dunn, J., and Kalish, M. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspect. Psychol. Sci.* 16, 767–778. doi: 10.1177/1745691620974771

Laming, D. (1991). Reconciling Fechner and Stevens? *Behav. Brain Sci.* 14, 188–191. doi: 10.1017/S0140525X0006605X

Laming, D. (2010). Fechner's law: Where does the log transform come from? *Seeing Perceiving* 23, 155–171. doi: 10.1163/187847510X503579

Luce, R. (1972). What sort of measurement is psychophysical measurement? *Am. Psychol.* 27, 96–106. doi: 10.1037/h0032677

Luce, R. D., and Edwards, W. (1958). The derivation of subjective scales from just noticeable differences. *Psychol. Rev.* 65, 222–237. doi: 10.1037/h0039821

Luchetti, M. (2020). From successful measurement to the birth of a law: Disentangling coordination in Ohm's scientific practice. *Stud. Hist. Philos. Sci.* 84, 119–131. doi: 10.1016/j.shpsa.2020.09.005

Luchetti, M. (2022). The quantification of intelligence in nineteenth-century craniology: An epistemology of measurement perspective. *Eur. J. Philos. Sci.* 12:56. doi: 10.1007/s13194-022-00485-7

Lundmann, L., and Villadsen, J. W. (2016). Qualitative variations in personality inventories: Subjective understandings of items in a personality inventory. *Qual. Res. Psychol.* 13, 166-187. doi: 10.1080/14780887.2015.1134737

Mach, E. (1872/1909). *Die Geschichte und die Wurzel des Satzes yon der Erhaltung der Arbeit, Calve, Prague, zweiter, unveränderter Abdruck*. Leipzig: J. A. Barth, 1909.

Mach, E. (1896/1986). *Principles of the theory of heat*. T.J. McCormack (trans.). Dordrecht: D. Reidel.

Marchionni, C., Zahle, J., and Godman, M. (2024). Reactivity in the human sciences. *Eur. J. Philos. Sci.* 14, 1–24. doi: 10.1007/s13194-024-00571-y

Mari, L. (2003). Epistemology of measurement. *Measurement* 34, 17–30. doi: 10.1016/S0263-2241(03)00016-2

Mari, L. (2013). A quest for the definition of measurement. *Measurement* 46, 2889–2895. doi: 10.1016/j.measurement.2013.04.039

Mari, L., Maul, A., Irribarra, D. T., and Wilson, M. (2016). A meta-structural understanding of measurement. *J. Phys.* 772:012009. doi: 10.1088/1742-6596/772/1/012009

Mari, L., Wilson, M., and Maul, A. (2023). *Measurement across the Sciences: Developing a Shared Concept System for Measurement*. New York, NY: Springer. doi: 10.1007/978-3-031-22448-5

Markus, K. A., and Borsboom, D. (2013). *). Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. London: Routledge. doi: 10.4324/9780203501207

Marshall, M. E. (1982). "Physics, metaphysics, and Fechner's psychophysics," in *The Problematic Science: Psychology in Nineteenth-Century Thought*, eds W. R. Woodward and M. G. Ash (New York, NY: Praeger Publishers), 65–87.

Marshall, M. E. (1990). The theme of quantification and the hidden Weber in the early work of Gustav Theodor Fechner. *Can. Psychol.* 31, 45–53. doi: 10.1037/h0078926

McClimans, L. (2013). The role of measurement in establishing evidence. *J. Med. Philos.* 38, 520–538. doi: 10.1093/jmp/jht041

McClimans, L. (2023). "Measurement, hermeneutics and standardization: why Gadamerian hermeneutics is necessary to contemporary philosophy of science," in *Updating the Interpretive Turn: New Arguments in Hermeneutics*, ed. M. Meijer (London: Routledge), 157–172. doi: 10.4324/9781003251354-12

McClimans, L., Browne, J., and Cano, S. (2017). Clinical outcome measurement: Models, theory, psychometrics and practice. *Stud. Hist. Philos. Sci. Part A* 65, 67–73. doi: 10.1016/j.shpsa.2017.06.004

McGrane, J. A. (2015). Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Front. Psychol.* 6:431. doi: 10.3389/fpsyg.2015.00431

Meischner-Metge, A. (2010). Gustav Theodor Fechner: life and work in the mirror of his diary. *Hist. Psychol.* 13, 411–423. doi: 10.1037/a0021587

Melin, J., Cano, S. J., Flöel, A., Göschel, L., and Pendrill, L. R. (2021). Construct specification equations: 'Recipes' for certified reference materials in cognitive measurement. *Measurement Sensors* 18:100290. doi: 10.1016/j.measen.2021.100290

Messick, S. (1989). "Validity," in *Educational Measurement*, ed. R. L. Linn (New York, NY: American Council on Education and Macmillan), 13–103.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *Br. J. Psychol.* 88, 355–383. doi: 10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in Psychology. A Critical History of a Methodological Concept*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511490040

Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometricians' fallacy. *Stud. Hist. Philos. Biol. Biomed. Sci.* 37, 414–432. doi: 10.1016/j.shpsc.2006.06.011

Michell, J. (2008). Is psychometrics pathological science? *Measurement* 6, 7–24. doi: 10.1080/15366360802035489

Michell, J. (2012). "The constantly recurring argument": Inferring quantity from order. *Theory Psychol.* 22, 255–271. doi: 10.1177/09593543114346

Murray, D. J. (1993). A perspective for viewing the history of psychophysics. *Behav. Brain Sci.* 16, 115–137. doi: 10.1017/S0140525X00029277

Muthukrishna, M., and Henrich, J. (2019). A problem in theory. *Nat. Hum. Behav.* 3, 221–229. doi: 10.1038/s41562-018-0522-1

Newton, P. E., and Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment*. London: Sage. doi: 10.4135/9781446288856

Nicolas, S. (2002). La fondation de la psychophysique de Fechner: Des présupposés métaphysiques aux écrits scientifiques de Weber. *L'Année Psychol.* 102, 255–298. doi: 10.3406/psy.2002.29592

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* 26, 1596–1618. doi: 10.3758/s13423-019-01645-2

Orne, M. Y. (1962). On the social psychology of the psychological experiment: With Particular reference to demand characteristics and their implications. *Am. Psychol.* 17, 776–783. doi: 10.1037/h0043424

Padovani, F. (2017). "Coordination and measurement: What we get wrong about what Reichenbach got right," in *EPSA15 Selected Papers. European Studies in Philosophy of Science*, Vol. 5, eds M. Massimi, J. W. Romeijn, and G. Schurz (Cham: Springer International Publishing), 49–60. doi: 10.1007/978-3-319-53730-6_5

Pendrill, L. (2019). *Quality Assured Measurement*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-28695-8

Pendrill, L., and Petersson, N. (2016). Metrology of human-based and other qualitative measurements. *Meas. Sci. Technol.* 27:094003. doi: 10.1088/0957-0233/27/9/094003

Reichenbach, H. (1920). *Relativitätstheorie und Erkenntnis Apriori*. Berlin: Springer. doi: 10.1007/978-3-642-50774-8

Riordan, S. (2015). The objectivity of scientific measures. *Stud. Hist. Philos. Sci. Part A* 50, 38–47. doi: 10.1016/j.shpsa.2014.09.005

Ryckman, T. A. (1991). Conditio sine qua non? Zuordnung in the early epistemologies of Cassirer and Schlick. *Synthese* 88, 57–95. doi: 10.1007/BF00540093

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Stud. Hist. Philos. Sci. Part A* 42, 509–524. doi: 10.1016/j.shpsa.2011.07.001

Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. London: Palgrave Macmillan. doi: 10.1057/978-1-137-38523-9

Slaney, K. L., and Garcia, D. A. (2015). Constructing psychological objects: The rhetoric of constructs. *J. Theor. Philos. Psychol.* 35, 244–259. doi: 10.1037/teo0000025

Smedslund, J. (2016). Why psychology cannot be an empirical science. *Integr. Psychol. Behav. Sci.* 50, 185–195. doi: 10.1007/s12124-015-9339-x

Staley, R. (2021). Sensory studies, or when physics was psychophysics: Ernst Mach and physics between physiology and psychology, 1860-71. *Hist. Sci.* 59, 93–118. doi: 10.1177/0073275318784104

Stevens, S. (1956). The direct estimation of sensory magnitudes-loudness. *Am. J. Psychol.* 69, 1–25. doi: 10.2307/1418112

Stevens, S. (1957). On the psychophysical law. *Psychol. Rev.* 64, 153–181. doi: 10.1037/h0046162

Stevens, S. (1969). Sensory scales of taste intensity. *Percept. Psychophys.* 6, 302–308. doi: 10.3758/BF03210101

Stevens, S. (1970). Neural events and the psychophysical law. *Science* 170, 1043–1050. doi: 10.1126/science.170.3962.1043

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Belknap Press of Harvard University Press.

Stone, C. (2019). A defense and definition of construct validity in psychology. *Philos. Sci.* 86, 1250–1261. doi: 10.1086/705567

Tal, E. (2017). Calibration: Modelling the measurement process. *Stud. Hist. Philos. Sci.* 65-66, 33–45. doi: 10.1016/j.shpsa.2017.09.001

Tal, E. (2019). Individuating quantities. *Philosophical Studies* 176, 853–878. doi: 10.1007/s11098-018-1216-2

Tal, E. (2020). Measurement in science. *The Stanford Encyclopedia of philosophy* (Fall 2020 Edition), ed. E. N. Zalta. Available online at: https://plato.stanford.edu/archives/fall2020/entries/measurement-science/

Tal, E. (2021). Two myths of representational measurement. *Perspect. Sci.* 29, 701–741. doi: 10.1162/posc_a_00391

Tannery, J. (1875a). Correspondence. A propos du logarithme des sensations. *La Revue Sci.* 15, 876–877.

Tannery, J. (1875b). La mesure des sensations: Réponses à propos du logarithme des sensations. *La Revue Sci.* 8, 1018–1020.

Teller, P. (2013). The concept of measurement-precision. *Synthese* 190, 189–202. doi: 10.1007/s11229-012-0141-8

Teller, P. (2018). "Measurement accuracy realism," in *The Experimental Side of Modeling*, eds I. Peschard and B. C. van Fraassen (Minneapolis: University of Minnesota Press), 273–298. doi: 10.5749/j.ctv5cg8vk.15

Thompson, M. (2023). Path-dependence in measurement: A problem for coherentism. *Philos. Sci.* 1, 1–11. doi: 10.1017/psa.2023.147

Toomela, A. (2008). Variables in psychology: A critique of quantitative psychology. *Integr. Psychol. Behav. Sci.* 42, 245–265. doi: 10.1007/s12124-008-9059-6

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory Psychol.* 19, 579–599. doi: 10.1177/0959354309341926

Uher, J. (2019). Data generation methods across the empirical sciences: Differences in the study phenomena's accessibility and the processes of data encoding. *Qual. Quant.* 53, 221–246. doi: 10.1007/s11135-018-0744-3

Uher, J. (2020). Measurement in metrology, psychology and social sciences: Data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Qual. Quant.* 54, 975–1004. doi: 10.1007/s11135-020-00970-2

Uher, J. (2021a). Quantitative psychology under scrutiny: Measurement requires not result-dependent but traceable data generation. *Pers. Individ. Differ.* 170:110205. doi: 10.1016/j.paid.2020.110205

Uher, J. (2021b). Psychology's Status as a Science: Peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integr. Psychol. Behav. Sci.* 55, 212–224. doi: 10.1007/s12124-020-09545-0

Uher, J. (2022). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: A rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780199278220.001.0001

von Kries, J. (1882). Ueber die Messung intensiver Grossen und uber das sogenannte psychophysische Gesetz. *Vierteljahrsschrift fur wissenschaftliche Philosophie* 6, 257-94. [Translation 1995 by K. K. Niall as "Conventions of Measurement in Psychophysics. Von Kries on the So-called Psychophysical Law." *Spatial Vision* 9, 275-305].

Wajnerman-Paz, A., and Rojas-Líbano, D. (2022). On the role of contextual factors in cognitive neuroscience experiments: a mechanistic approach. *Synthese* 200:402. doi: 10.1007/s11229-022-03870-0

Wasserman, G. S., Felsten, G., and Easland, G. S. (1979). The psychophysical function: Harmonizing Fechner and Stevens. *Science* 204, 85–87. doi: 10.1126/science.432630

Weber, E. H. (1834). *De Pulsu, Resorptione, Auditu et tactu: Annotationes Anatomicae et Physiologicae*. Leipzig: C. F. Koehler.

Weber, E. H. (1846). "Der Tastsinn und das Gemeingefuhl," in *Handworterbuch der Physiologie, mit Rucksicht auf physiologische Pathologie*, Vol. 3, ed. R. Wagner (Braunschweig: Vieweg), 481–588.

Zhao, K. (2023). Measuring the nonexistent: Validity before measurement. *Philos. Sci.* 90, 227–244. doi: 10.1017/psa.2023.3

Zudini, V. (2011). The Euclidean model of measurement in Fechner's psychophysics. *J. Hist. Behav. Sci.* 47, 70–87. doi: 10.1002/jhbs.20472

frontiers | Frontiers in Psychology

TYPE Perspective
PUBLISHED 30 May 2024
DOI 10.3389/fpsyg.2024.1383622



OPEN ACCESS

EDITED BY
Barbara Hanfstingl,
University of Klagenfurt, Austria

REVIEWED BY
Matthias Borgstede,
University of Bamberg, Germany
Michele Luchetti,
Bielefeld University, Germany

*CORRESPONDENCE
Josh Joseph Ramminger
✉ josh.ramminger@hu-berlin.de
Niklas Jacobs
✉ jacobnik@hu-berlin.de

RECEIVED 07 February 2024
ACCEPTED 03 May 2024
PUBLISHED 30 May 2024

CITATION
Ramminger JJ and Jacobs N (2024) Primacy of theory? Exploring perspectives on validity in conceptual psychometrics.
*Front. Psychol.* 15:1383622.
doi: 10.3389/fpsyg.2024.1383622



# Primacy of theory? Exploring perspectives on validity in conceptual psychometrics

Josh Joseph Ramminger[1,2]* and Niklas Jacobs[1]*

[1]Faculty of Life Sciences, Department of Psychology, Humboldt University of Berlin, Berlin, Germany, [2]Department of Psychology, Philipps-University of Marburg, Marburg, Germany

Several conceptions of validity have emphasized the contingency of validity on theory. Here we revisit several contributions to the discourse on the concept of validity, which we consider particularly influential or insightful. Despite differences in metatheory, both Cronbach and Meehl's construct validity, and Borsboom, Mellenbergh and van Heerden's early concept of validity regard validity as a criterion for successful measurement and thus, as crucial for the soundness of psychological science. Others, such as Borgstede and Eggert, regard recourses to validity as an appeal to an (unscientific) folk psychology. Instead, they advocate theory-based measurement. It will be demonstrated that these divergent positions converge in their view of psychological theory as indispensable for the soundness of psychological measurement. However, the formulation of the concept (and scope) of scientific theory differs across the presented conceptions of validity. These differences can be at least partially attributed to three disparities in metatheoretical and methodological stances. The first concerns the question of the structure of scientific theories. The second concerns the question of psychology's subject matter. The third regards whether, and if, to which extent, correlations can be indicative of causality and therefore point toward validity. These results indicate that metatheory may help to structure the discourse on the concept of validity by revealing the contingencies the concrete positions rely on.

KEYWORDS

**validity, theory, conceptual psychometrics, philosophy of science, metatheory, methodology**

## Introduction

How shall we understand the concept of validity? Which methodological implications arise from conceptions and critique of validity? These questions have been subject to a lively discourse. Within this discourse, substantial divergence regarding metatheory and methodology in psychology is present (see Cronbach and Meehl, 1955; Messick, 1989; Slaney, 2017; Borsboom, 2023). For us, metatheory deals with the investigation of scientific theories, as well as their relation to stances in theory of science. In our view, philosophy and the sciences can particularly benefit from the investigation of the logical connection between metatheory and methodology (see Hanfstingl, 2019; Uher, 2023). Validity, as one domain of disagreement, is commonly understood to address whether one measures what is intended to be measured. However, this definition has been criticized because it presupposes that one *is* measuring something and that that which shall be measured is measurable (Michell, 2009, 11–33). For some validity concerns the soundness of a conclusion drawn from a measurement outcome (see Markus and Borsboom, 2013). One of us has argued elsewhere that the validity debate is

a prime example of a philosophical-psychological discourse, as in it logical connections between metatheory and methodology are illustrated (Ramminger, 2023).

One such logical connection is *scientific theory*. Philosophy of science investigates the structure of scientific theories (e.g., Balzer et al., 1987). Metatheoretical assumptions can structure scientific theories because scientific theories can deal with the same entities based on the same empirical evidence and still be different (Ramminger et al., 2023). Concrete (i.e., clearly defined) scientific theories are furthermore an important element of the working scientists' epistemic processes (Hastings et al., 2020).

However, scientometric studies show that not all psychological research can be regarded as theory-driven (McPhetres et al., 2021; Wendt and Wolfradt, 2022), even though low replication rates in psychology have repeatedly been attributed to deficiencies in theory-building and application (Fiedler, 2017; Muthukrishna and Henrich, 2019; Oberauer and Lewandowsky, 2019; Green, 2021; Witte, 2022; Ramminger et al., 2023). Such agreement lacks regarding the relationship between theory and validity, even though validity is a prerequisite for replicability (Flake et al., 2022). For example, Borgstede (2019) has argued that some applied validity research is atheoretical. In addition, different theory-based conceptions of validity differ in their concept of scientific theories (Borsboom et al., 2004; Buntins et al., 2017). Furthermore, even when adhering to one specified conception of validity (such as Cronbach and Meehl's construct validity), the underlying theory (i.e., the nomological net) is not always stated explicitly (for an introduction see Ziegler et al., 2013).

In what follows, we will show that different conceptions of validity and validity's relation to scientific theory stem from metatheoretical assumptions. These differences concern the structure of scientific theories, the question of psychology's subject matter (Wendt and Funke, 2022; Wendler and Ramminger, 2023), as well as methodological considerations (e.g., *whether, and if, to which extent* correlations can be indicative for causality and therefore pointing toward validity). Finally, we will show that proponents and critics of the employment of validity converge in their assumption that *theory-basedness* is at least necessary to ensure the soundness of psychological measurement.

## Metatheory, validity, and scientific theory

Several conceptions of validity can be traced back to their metatheoretical assumptions. Some movements in philosophy of science have therefore been associated with conceptions of validity. Examples range from descriptive empiricism, in the case of criterion validity, to logical positivism and scientific realism, in the case of construct validity (see Markus and Borsboom, 2013, 5–14; Slaney, 2017).[1] Furthermore, the semantic view of scientific theories (Balzer et al., 1987) is part of Borgstede and Eggert's account of theory-based measurement (Borgstede and Eggert, 2023). Our aim is not to settle questions in philosophy of science, but to demonstrate that different conceptions of validity converge in their assumption that validity is contingent upon theory. Moreover that this convergence of positions is present despite the divergent philosophies of science to which the positions adhere.

Different metatheoretical assumptions commonly entail a view on the nature of psychological attributes. Psychometricians often conceptualize their object of measurement as an unobservable mental construct (and consistently apply latent variable modeling). However, Borgstede and Eggert (2023) tend toward a behaviorist's perspective, thus seeing behavior as the crucial subject matter of psychology. Borsboom (2023) speaks of psychological attributes as *organizing principles* and thus adheres to network psychometrics and advocates for the rehabilitation of content validity. We are concerned here with the question of how authors of different perspectives in theory of science approach the relationship between validity and theory. We will present three positions associated with individual authors more in-depth, Cronbach and Meehl (1955), Borsboom's early perspective (Borsboom et al., 2004), and Borgstede and Eggert's (2023) position which rejects the term validity altogether but is still concerned with ensuring that psychologists know what they measure.

These accounts were selected due to several factors. Cronbach and Meehl (1955) developed construct validity, arguably the conception of validity most utilized in contemporary psychology, for example it is largely adopted by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Consequently, the other accounts engage with Cronbach and Meehl (1955), while Borgstede and Eggert (2023) also address Borsboom et al. (2004). Furthermore, the term validity either denotes a characteristic of tests or test score interpretations (Borsboom et al., 2003b; Borsboom and Markus, 2013). Two selected papers (Borsboom et al., 2004; Borgstede and Eggert, 2023) engage with the first meaning, while Cronbach and Meehl (1955) aim to address the second one. The selected stances diverge in philosophical questions (e.g., realism or how scientific theories shall be structured), however one can logically infer from these approaches, that validity must be theory based. This convergence—despite philosophical divergence - is thus a strong argument for the necessity of theory for validity. Lastly, Borgstede and Eggert (2023) developed their approach analogous to measurement and theory building in the natural sciences whose methodological rigor is an ideal often adhered to in psychology (see James, 1892; Wieczorek et al., 2021).

First, we turn to *construct validity* (Cronbach and Meehl, 1955). As noted above, construct validity was associated with several traditions in philosophy. Since we aim to demonstrate that several accounts of validity are influenced by metatheoretical stances, more specifically traditions in philosophy of science and that these accounts align in their emphasis on the importance of scientific theory, we do not settle the question whether construct validity is indeed contingent to logical positivism (as Borsboom et al., 2004 argue) or scientific realism

1 Some scholars argue that logical positivism and empiricism should be used as a synonym (*cf.* Uebel, 2013). Markus and Borsboom (2013, 5–14) distinguish different forms of empiricism and relate these with different approaches to validity. We stick with their taxonomy for the purpose of differentiating between different metatheoretical foundations for validity concepts. We wish to thank the reviewers Matthias Borgstede and Michele Luchetti for providing valuable feedback, for example pointing towards this argument

(Rozeboom, 1984; Slaney, 2012; see also Slaney, 2017). However, since two of the three positions we address *definitely* reject logical positivism (see Borsboom et al., 2004; Borgstede, 2022), we focus here on a logical positivist's interpretation of construct validity (for an introduction to logical positivism see Creath, 2023) to stretch the logical space and show that validity conceptions resting on logical positivism likewise regard a well-formulated theory as a prerequisite for the investigation of a measurement instrument's validity.[2]

Such an account of construct validity emphasises that (a) Cronbach and Meehl insisted that the nomological network gives constructs their meaning (by making the relations of the constructs explicit) and that (b) Cronbach and Meehl are especially concerned with cases in which at least one variable studied cannot be regarded as observable, i.e., they are interested in the relation of theoretical constructs to observables. For example, if you were to create a conscientiousness personality test item based on this account, you would *a priori* point out expected relations (i.e., a high correlation with average punctuality). After a first test phase of the item, you would either confirm this expectation, concluding that you measured conscientiousness or, in case you found an unexpected correlation, conclude that you did not measure conscientiousness/create a new hypothesis that conscientiousness in fact does not correlate highly with punctuality (see Cronbach and Meehl, 1955).

Consistently, according to Borgstede, the positivist assumes the task of science to be translating observations into theory-language to determine the truth of the theoretical propositions. This practice would be contingent on a syntactic conception of scientific theories. The syntactic view regards a scientific theory as a system of propositions. These syntactic structures are identified by applying the theory to empirical relational structures through operationalization, or correspondence rules as they are called in theory of science (Borgstede, 2022, 18–19). Therefore, the relation between observables and non-observables is a central element of construct validity *and* positivism.

The importance of scientific theory in determining construct validity can be further demonstrated by Cronbach and Meehl's assertion that the "types of evidence" for construct validity depend "on the theory surrounding the construct" (Cronbach and Meehl, 1955, 288). Such types of evidence could be factor analyses, another one correlations. Moreover, the execution of a measurement may result in two potential outcomes: either concluding that the results indicate construct validity or adjustment of the nomological net, consequently impacting the underlying theory. Thus, construct validity is judged after measurement.

Borsboom and several colleagues, the second position we review more in-depth, disagree with the metatheoretical stances of a positivist's reading of construct validity. In their early work, Borsboom and several colleagues advocate for a validity concept based on a realist's metatheory (Borsboom et al., 2004, 2009). For Borsboom, Mellenbergh, and van Heerden, logical positivism and its application to validity theory rests on the possibility of making meaningful statements without referring to existing attributes.

For the logical positivist, advocating for construct validity, a test could be regarded as valid for measuring a construct if the empirical relations between test scores match the theoretical relations between constructs. That theorist would continue to argue that the meaning of psychological constructs is determined via the relation of the corresponding concepts in a nomological network. In contrast, Borsboom et al. adhere to a realist account of validity, since they regard it as inconceivable, "how the sentences *Test X measures the attitude toward nuclear energy* and *Attitudes do not exist* can both be true" (Borsboom et al., 2004, 1063). Their commitments to philosophical realism (see also Borsboom et al., 2003a; Borsboom, 2005, 6–8; Borsboom also quotes Hacking, 1983 and Devitt, 1991 when introducing realism) allow Borsboom and colleagues to infer two crucial methodological implications. First, they regard a test as being "valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al., 2004, 1061). Secondly, a theory about the response behavior of people is necessary, otherwise, validity judgements cannot be made. In other words, if the attribute causes variations in the test scores, this causal influence must occur somewhere in the process of responding itself and theories have to take this response process into account.[3]

To better understand this approach, we can again refer to our example of the conscientiousness item. Following Borsboom's and colleagues' 2004 approach, you would establish a theory of the causal role of conscientiousness for the response given to the item. For example, conscientious people will read the item carefully and unveil an ambiguity, which evokes an answer divergent from non-conscientious people.

How can one test this theory? One could infer that the answers given by divergent subgroups which are expected to be very conscientious (potentially air traffic controllers) differ from the answers given by groups that are expected to be less conscientious (potentially graphic designers, this example only has illustrative purpose). Note that this represents a test of the underlying theory, not the validity of the conscientiousness item.

The question of validity thus becomes the question whether the attribute of interest exists and how that attribute—this is where the theory comes in—causally affects test scores.

Furthermore, Borsboom et al. (2004) criticize correlation-based and anti-realist positions approaches to validity, since two absurd conclusions would follow from them. Firstly that two highly correlated constructs are identical (see also Borgstede's (2019) critique), and secondly that when measuring a group of objects that do not show variation in the interesting attribute, it would become *a priori* impossible to conclude that the measurement is valid since for a variance of zero the correlation is undefined.[4] Suppose one wants to

2 However, it must be noted that since 1955 construct validity has evolved and that Cronbach in his later work regarded it as problematic to formulate the idea of construct validity in the language of positivism (Cronbach, 1989, 159; see also Slaney, 2017).

3 It is not our intention to assert that Borsboom and the several colleagues, with which he put forth this conception still adhere to this position. As we have briefly touched upon, Borsboom recently elaborated on validity in network psychometrics and the implications of this approach for the ontology of psychological attributes (Borsboom, 2023). However, the early work with which we engage here is customized for latent variable analysis, which is still widely applied in psychometrics.

4 This can be derived from $r = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$ since one variance being zero ($x$ or $y$) leads to a division by zero, which is undefined.

measure the length of rods using a meter stick and that all rods have the same length. One could not conclude that the meter stick is a valid measure of length (see Borsboom et al., 2004).

Finally, Borsboom et al. (2004) emphasis on ontology in validity leads them to critique positions that regard validity to be judged *after measurement,* since knowledge of the nature of the object of measurement would imply knowledge of the steps one has to take to measure that object. Thus, validity would become an *a priori* matter of metatheory (ontology) and scientific theory. Ontology deals with condition (a), the existence of the attribute, and scientific theory with condition (b), whether 'variations in the attribute causally produce variations in the outcomes of the measurement procedure'.

As a third perspective, Borgstede and several colleagues do not agree that the attribute necessarily exists (Buntins et al., 2017). They claim that the central problem of psychological measurement is not unobservability, but the lack of well-defined concepts. Like Borsboom and colleagues, they explicitly reject logical positivism and the associated *syntactic* view of scientific theories (Borgstede and Eggert, 2023). Borgstede and Eggert follow the *semantic* view in theory of science, according to which a *substantive* fundamental principle structures a scientific theory (Borgstede, 2022; Borgstede and Eggert, 2023; for an philosophical introduction see Balzer et al., 1987). For Borgstede, one such principle might be behavioral selection (Borgstede, 2022, 31). Since behavior is observable, the problem of psychological concepts for Borgstede and Eggert is not that they are observable or latent, but that they are *poorly defined* (Borgstede and Eggert, 2023).

According to Borgstede and Eggert, one cannot determine whether one measures what one wants to measure independently of a measurement theory. When using the operational theory of measurement (see Stevens, 1946) one (by definition) measures what one intends to measure, since there is no difference between what is to be measured and the indicator. The representational measurement theory (see Krantz et al., 1971), however, gives testable criteria for investigating whether one is measuring what one wants to measure (Buntins et al., 2017).

Consistently, for Borgstede and Eggert, the problem with psychological concepts is that they are rarely defined within the framework of a substantive (formal) theory (Borgstede and Eggert, 2023). In this context, a substantive (and) formal theory can be described as a hierarchical network. Substantively, this network is structured by a *fundamental* underlying principle (e.g., behavioral selection) and more *specific* principles (e.g., specific types of reinforcement) that explain empirical phenomena (e.g., change in behavior). These principles are formally defined (Borgstede and Eggert, 2023). This often implies a mathematical definition, but one can also find formalizations that utilize formal logic (Buntins et al., 2015). In psychology, descriptively speaking, validity would commonly term "the degree to which the variable measured by a test corresponds to concepts of everyday language" (Buntins et al., 2017). However, if validity is supposed to anchor psychological concepts in common-sense, which trivially is not mathematically accurate, then it is not possible to measure in a theory-based way.

Their proposed antidote is theory-based measurement, which they regard as necessary and sufficient for *knowing what we are measuring*. That is because proper theory informs us about the steps necessary to measure the entities the theory entails. Put differently, the knowledge of the measurement procedure stems from the theorized relation between the objects of measurement (observable phenomena). For example, we can adhere to Newton's second law to measure mass, since it allows us to use a beam scale (Borgstede and Eggert, 2023, for an application to behavioral measurement see Borgstede and Anselme, 2024).

How would this approach relate to our running example of the construction of a conscientiousness item? This is a puzzling question, and one could even argue that here we face the danger of a category error. It is tempting to understand conscientiousness as a mental attribute, in which case it would not be straightforward to align the concept of conscientiousness with Borgstede's behaviorist leanings.

Borgstede suggests behavioral selection as a fundamental principle which can structure psychological theories. The content of these theories should be the interaction of individuals and their environment. Therefore, conscientiousness possibly needs a redefinition regarding its causal relation to the fundamental principle and the other entities postulated in the general theory. Drawing on Borgstede's exemplary fundamental principle of behavioral selection, one would have to relate conscientiousness to it and the less abstract entities and principles in the theory net. Such a relation could draw on principles of social interaction in early human societies, which could potentially contribute to the explanation of the genesis of conscientiousness from natural selection. Another possibility is that such a theory would not include entities that correspond to concepts that are derived from common language. In this case, one may conclude that conscientiousness does not exist.

Comparing the three discussed accounts of validity and their relation to theory, several aspects deserve additional emphasis. Although Borgstede and Eggert reject the recourse to validity in the sense the term is often used in psychology, they still regard theory as necessary to solve the epistemic questions the validity discourse raises. Logically, Borsboom et al. (2004) are concerned with something akin, since they reject the idea that one can determine whether one has measured what one wanted to measure *after* the measurement procedure (unlike Cronbach and Meehl). Since theories describe causal processes, Borsboom et al. (2004), as well as Borgstede and Eggert (2023), Borgstede (2019) converge in the assumption that determining validity implies that we need to adhere to an *a priori* theory of the *causal* properties of our variable of interest. Thus, they stand in stark opposition to Cronbach and Meehl's approach of judging construct validity *a posteriori* (possibly based on correlations, which are viewed as indicative of causality). All three positions presented formulate their idea of psychological measurement, to which its construct is known, within the context of a philosophy of science and attribute central relevance to scientific theories.

## Conclusion and limitations

All three positions align in emphasizing the central relevance of scientific theory in understanding and defining validity in psychological measurement. They all underscore the importance of having a well-formulated theoretical framework when considering the validity of measurement instruments in psychology. However, they differ in their specific philosophical and metatheoretical assumptions, as well as in the question of whether validity is judged

*a priori* or *a posteriori*. Consistently, two of the formulated approaches reject the inference of validity from empirical results (e.g., correlation matrices), since they adhere to measurement procedures derived by *a priori* reflections on the causal properties of the variable under investigation. They thus emphasize that validity conclusions are justified by adherence to theoretical propositions. Of course, the quality of validity concepts depends on metatheoretical criteria such as consistency. Furthermore, the question of the feasibility of the methodological implications in research projects is also highly relevant (see also Borsboom, 2023). However, questions about the criteria of measurability (e.g., Michell, 1999; Markus and Borsboom, 2012) and the potential context dependency of validity (see for a critique, Larroulet Philippi, 2021) exceed the scope of this paper. After all, in this essay, we were concerned precisely with the inner, logical, relationship of metatheory and methodology in the discourse on validity. Logically, all three positions engage with the conditions of knowing what psychologists are measuring (*a priori* or *a posteriori*), therefore Borgstede and Eggert are part of this discourse, even though they reject the term (and a certain notion of) validity. This paper demonstrates the interconnectedness of metatheory, theory, and measurement and aims to encourage an appreciation of theory for the soundness of psychological measurement, which is not always present in contemporary psychometrics.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

## Author contributions

JR: Conceptualization, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. NJ: Conceptualization, Funding acquisition, Investigation, Methodology, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The article processing charge was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 491192747 and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Balzer, W., Moulines, C. U., and Sneed, J. D. (1987). *An architectonic for science. The structuralist program*. Dordrecht: Reidel.

Borgstede, M. (2019). Zwischen Definition und Empirie. *Vierteljahresschrift für wissenschaftliche Pädagogik* 95, 199–217. doi: 10.30965/25890581-09501018

Borgstede, M. (2022). *Theorie und Messung in der Psychologie: eine evolutionäre Perspektive*. Bamberg: Schriften aus der Fakultät Humanwissenschaften der Otto-Friedrich-Universität Bamberg.

Borgstede, M., and Anselme, P. (2024). Model-based estimates for operant selection. *bioRxiv preprints*. doi: 10.1101/2022.07.22.501082

Borgstede, M., and Eggert, F. (2023). Squaring the circle: from latent variables to theory based measurement. *Theory Psychol.* 33, 118–137. doi: 10.1177/09593543221127985

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2023). "Psychological constructs as organizing principles" in *Essays on Contemporary Psychometrics*. eds. L. A. van der Ark, W. H. M. Emons and R. R. Meijer (Cham: Springer), 89–108.

Borsboom, D., and Markus, K. A. (2013). Truth and evidence in validity theory. *J. Educ. Meas.* 50, 110–114. doi: 10.1111/jedm.12006

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2003a). The theoretical status of latent variables. *Psychol. Rev.* 110, 203–219. doi: 10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., and Van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Borsboom, D., Van Heerden, J., and Mellenbergh, G. J. (2003b). "Validity and truth" in *New developments in psychometrics*. eds. H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J. J. Meulman (Tokyo: Springer Japan), S. 321–S. 328.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., and Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), The concept of validity. 135–170. Charlotte, NC: Information Age

Buntins, M., Buntins, K., and Eggert, F. (2015). Psychological tests from a (fuzzy-) logical point of view. *Qual. Quant.* 50, 2395–2416. doi: 10.1007/s11135-015-0268-z

Buntins, M., Buntins, K., and Eggert, F. (2017). Clarifying the concept of validity: from measurement to everyday language. *Theory Psychol.* 27, 703–710. doi: 10.1177/0959354317702256

Creath, R. (2023). "Logical empiricism" in *The Stanford encyclopedia of philosophy*. eds. E. N. Zalta and U. Nodelman. Winter 2023 ed. https://plato.stanford.edu/archives/win2023/entries/logical-empiricism/.

Cronbach, L. J. (1989). "Construct validation after thirty years" in *Intelligence: Measurement, theory, and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys*. ed. R. L. Linn (University of Illinois Press), 147–171.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Devitt, M. (1991). *Realism and truth. 2nd* Edn: Blackwell.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspect. Psychol. Sci.* 12, 46–61. doi: 10.1177/1745691616654458

Flake, J. K., Davidson, I. J., Wong, O., and Pek, J. (2022). Construct validity and the validity of replication studies: a systematic review. *Am. Psychol.* 77, 576–588. doi: 10.1037/amp0001006

Green, C. D. (2021). Perhaps psychology's replication crisis is a theoretical crisis that is only masquerading as a statistical one. *Int. Rev. Theor. Psychol.* 1. doi: 10.7146/irtp.v1i2.127764

Hacking, I. (1983). Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge: Cambridge University Press.

Hanfstingl, B. (2019). Should we say goodbye to latent constructs to overcome replication crisis or should we take into account epistemological considerations? *Front. Psychol.* 10:1949. doi: 10.3389/fpsyg.2019.01949

Hastings, J., Michie, S., and Johnston, M. (2020). Theory and ontology in behavioural science. *Nat. Hum. Behav.* 4:226. doi: 10.1038/s41562-020-0826-9

James, W. (1892). A plea for psychology as a "natural science". *Philos. Rev.* 1, 146–153. doi: 10.2307/2175743

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971). *Foundations of measurement. Vol. I. Additive and polynomial representations*. San Diego, CA: Academic Press.

Larroulet Philippi, C. (2021). Valid for what? On the very idea of unconditional validity. *Philos. Soc. Sci.* 51, 151–175. doi: 10.1177/0048393120971169

Markus, K. A., and Borsboom, D. (2012). The cat came back: evaluating arguments against psychological measurement. *Theory Psychol.* 22, 452–466. doi: 10.1177/0959354310381155

Markus, K. A., and Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

McPhetres, J., Albayrak-Aydemir, N., Mendes, A. B., Chow, E. C., Gonzalez-Marquez, P., Loukras, E., et al. (2021). A decade of theory as reflected in psychological science (2009–2019). *PLoS One* 16:e0247986. doi: 10.1371/journal.pone.0247986

Messick, S. (1989). "Validity" in *Educational measurement*. ed. R. L. Linn (Washington, DC: American Council on Education and National Council on Measurement in Education), 13–103.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept, vol. 53*. Cambridge: Cambridge University Press.

Michell, J. (2009). "Invalidity in validity" in *The concept of validity: revisions, new directions and applications*. ed. R. W. Lissitz (IAP Information Age Publishing), 111–134.

Muthukrishna, M., and Henrich, J. (2019). A problem in theory. *Nat. Hum. Behav.* 3:221229, 221–229. doi: 10.1038/s41562-018-0522-1

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* 26, 1596–1618. doi: 10.3758/s13423-019-01645-2

Ramminger, J. J. (2023). Crossing the chasm? On the possibility of philosophical contributions to the discourse of quantitative psychology. *Cultura Psyché* 4, 215–224. doi: 10.1007/s43638-023-00081-3

Ramminger, J. J., Peper, M., and Wendt, A. N. (2023). Neuropsychological assessment methodology revisited: meta theoretical reflections. *Front. Psychol.* 14:1170283. doi: 10.3389/fpsyg.2023.1170283

Rozeboom, W. W. (1984). "Dispositions do explain: picking up the pieces after hurricane Walter" in *Annals of theoretical psychology (vol. 1)*. eds. J. R. Royce and L. P. Mos (New York, NY: Plenum), 205–224.

Slaney, K. L. (2012). Laying the cornerstone of construct validity theory: Herbert Feigl's influence on early specifications. *Theory Psychol.* 22, 290–309. doi: 10.1177/0959354311400659

Slaney, K. L. (2017). *Validating psychological constructs*: Palgrave Macmillan.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Uebel, T. (2013). "Logical positivism"—"logical empiricism": What's in a name? *Perspect. Sci.* 21, 58–99. doi: 10.1162/posc_a_00086

Uher, J. (2023). What are constructs? Ontological nature, epistemological challenges, theoretical foundations and key sources of misunderstandings and confusions. *Psychol. Inq.* 34, 280–290. doi: 10.1080/1047840x.2023.2274384

Wendler, H., and Ramminger, J. J. (2023). Was kann die phänomenologische Psychologie zur Verjüngung der Gegenstandsfrage beitragen? *J. Psychol.* 31, 59–81. doi: 10.30820/0942-2285-2023-1-59

Wendt, A. N., and Funke, J. (2022). *Wohin steuert die Psychologie?: Ein Ausrichtungsversuch (vol. 21)*: Vandenhoeck & Ruprecht.

Wendt, A. N., and Wolfradt, U. (2022). Theoretical psychology: discursive transformations and continuity in psychological research/Psychologische Forschung. *Psychol. Res.* 86, 2321–2340. doi: 10.1007/s00426-022-01727-2

Wieczorek, O., Unger, S., Riebling, J., Erhard, L., Koß, C., and Heiberger, R. (2021). Mapping the field of psychology: trends in research topics 1995–2015. *Scientometrics* 126, 9699–9731. doi: 10.1007/s11192-021-04069-9

Witte, E. H. (2022). Wissenschaftsgeschichte, Forscher_innengenerationen und die Vertrauenskrise in der Psychologie. *Psychol. Rundschau* 73, 41–42. doi: 10.1026/0033-3042/a000573

Ziegler, M., Booth, T., and Bensch, D. (2013). Getting entangled in the nomological net: thoughts on validity and conceptual overlap. *Eur. J. Psychol. Assess.* 29, 157–161. doi: 10.1027/1015-5759/a000173

frontiers | Frontiers in Psychology

TYPE Perspective
PUBLISHED 30 August 2024
DOI 10.3389/fpsyg.2024.1404060



OPEN ACCESS

EDITED BY
Mengcheng Wang,
Guangzhou University, China

REVIEWED BY
Longxi Li,
University of Washington, United States
Jie Luo,
Guizhou Normal University, China

*CORRESPONDENCE
Barbara Hanfstingl
barbara.hanfstingl@aau.at

RECEIVED 20 March 2024
ACCEPTED 14 August 2024
PUBLISHED 30 August 2024

CITATION
Hanfstingl B, Oberleiter S, Pietschnig J, Tran US and Voracek M (2024) Detecting jingle and jangle fallacies by identifying consistencies and variabilities in study specifications – a call for research. *Front. Psychol.* 15:1404060. doi: 10.3389/fpsyg.2024.1404060



# Detecting jingle and jangle fallacies by identifying consistencies and variabilities in study specifications – a call for research

Barbara Hanfstingl[1]*, Sandra Oberleiter[2], Jakob Pietschnig[2], Ulrich S. Tran[3] and Martin Voracek[3]

[1]Department of Psychology, University of Klagenfurt, Klagenfurt, Austria, [2]Department of Developmental and Educational Psychology, University of Vienna, Vienna, Austria, [3]Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Vienna, Austria

Over the past few years, more attention has been paid to jingle and jangle fallacies in psychological science. Jingle fallacies arise when two or more distinct psychological phenomena are erroneously labeled with the same term, while jangle fallacies occur when different terms are used to describe the same phenomenon. Jingle and jangle fallacies emerge due to the vague linkage between psychological theories and their practical implementation in empirical studies, compounded by variations in study designs, methodologies, and applying different statistical procedures' algorithms. Despite progress in organizing scientific findings via systematic reviews and meta-analyses, effective strategies to prevent these fallacies are still lacking. This paper explores the integration of several approaches with the potential to identify and mitigate jingle and jangle fallacies within psychological science. Essentially, organizing studies according to their specifications, which include theoretical background, methods, study designs, and results, alongside a combinatorial algorithm and flexible inclusion criteria, may indeed represent a feasible approach. A jingle-fallacy detector arises when identical specifications lead to disparate outcomes, whereas jangle-fallacy indicators could operate on the premise that varying specifications consistently yield overrandomly similar results. We discuss the role of advanced computational technologies, such as Natural Language Processing (NLP), in identifying these fallacies. In conclusion, addressing jingle and jangle fallacies requires a comprehensive approach that considers all levels and phases of psychological science.

KEYWORDS

**jingle fallacies, jangle fallacies, validity, meta-analysis, systematic review, specification analysis, harvest plot**

## Problem outline

In recent years, there has been increased attention on jingle and jangle fallacies in psychological science (Altgassen et al., 2024; Ayache et al., 2024; Beisly, 2023; Fischer et al., 2023; Hook et al., 2015; Marsh et al., 2019; Porter, 2023). Jingle fallacies occur when two or more distinct psychological phenomena are labeled with the same name, as Thorndike (1904,

p. 14) defined over 120 years ago. Kelley (1927, p. 64) later defined jangle fallacies as labeling the same phenomenon with different terms, exemplified by his use of 'intelligence' and 'achievement'. Gonzalez et al. (2021) highlighted that jingle and jangle fallacies pose a significant threat to the validity of the research. These fallacies are not always explicitly labeled as such; they may also be characterized as a déjà-variable phenomenon (Hagger, 2014; Hanfstingl, 2019; Skinner, 1996).

Why do jingle and jangle fallacies emerge? In essence, Thorndike (1904) and Kelley (1927) attributed their occurrence to a vague connection between psychological theory and its operationalization in empirical studies. Recent studies have emphasized the caution needed regarding jingle-jangle fallacies due to differences in algorithms used in statistical procedures (Grieder and Steiner, 2022). Another reason that exacerbates this problem is the substantial increase in scientific research since the Second World War, which has led to an increase in the overall number of studies carried out. However, as scientific knowledge continues to expand, there is an increasing need for its systematic organization and categorization. Without adequate systematization, the risk of poorly aligned parallel fields and trends operating independently increases, resulting in a disjointed theoretical landscape lacking overarching theories or paradigms. Finally, efficient progress is hindered by undetected inconsistencies in empirical evidence. Despite the long-standing knowledge of jingle and jangle fallacies, effective strategies to prevent psychological science from encountering these issues have not yet been developed.

In the 1970s, several solutions emerged to address the lack of systematization in scientific findings, with the development of review and meta-analytical approaches, albeit without explicit reference to jingle or jangle fallacies. According to Shadish and Lecy (2015), meta-analysis is considered "one of the most significant methodological advancements in science over the past century" (p. 246). Notably, Gene V. Glass focused on psychotherapy effects, Frank L. Schmidt emphasized psychological test validity, and Robert Rosenthal aimed to synthesize findings on interpersonal expectancy effects, all of whom contributed significantly to the development of meta-analysis (Shadish and Lecy, 2015).

While the practice of summarizing single studies in reviews and meta-analytical procedures has become common and well-accepted, several problems have become apparent: Systematic reviews and meta-analyses, while valuable, are not immune to bias and fail to detect jingle or jangle fallacies. Despite several initiatives like the PRISMA statement (Page et al., 2021) or meta-analysis reporting standards (MARS; Lakens et al., 2017), they still lack quality criteria (Glass, 2015; Pigott and Polanin, 2020) or ignore the influence of methodologies on the result (Elson, 2019). Some biases are extremely difficult to control, as, for example, those caused by scientists themselves (Hanfstingl, 2019; Wicherts et al., 2016) or by operationalization variances (Simonsohn et al., 2020; Steegen et al., 2016; Voracek et al., 2019). Furthermore, as with single studies, without transparency and free access to each point of the research process, reproducibility is not given (Maassen et al., 2020; Polanin et al., 2020). In sum, current review and meta-analytical approaches fail to uncover jingle or jangle fallacies.

## Approaches for detecting and preventing jingle and jangle fallacies

Essentially, we need not only programs to systematize empirical evidence and knowledge but also strategies to detect and prevent jingle and jangle fallacies, ideally combining single-study analyses at the meta-level. To address these challenges, we explore several potentially beneficial approaches. One such approach involves the systematization not only of results but also of theoretical backgrounds, methodological approaches, study designs, and outcomes. This

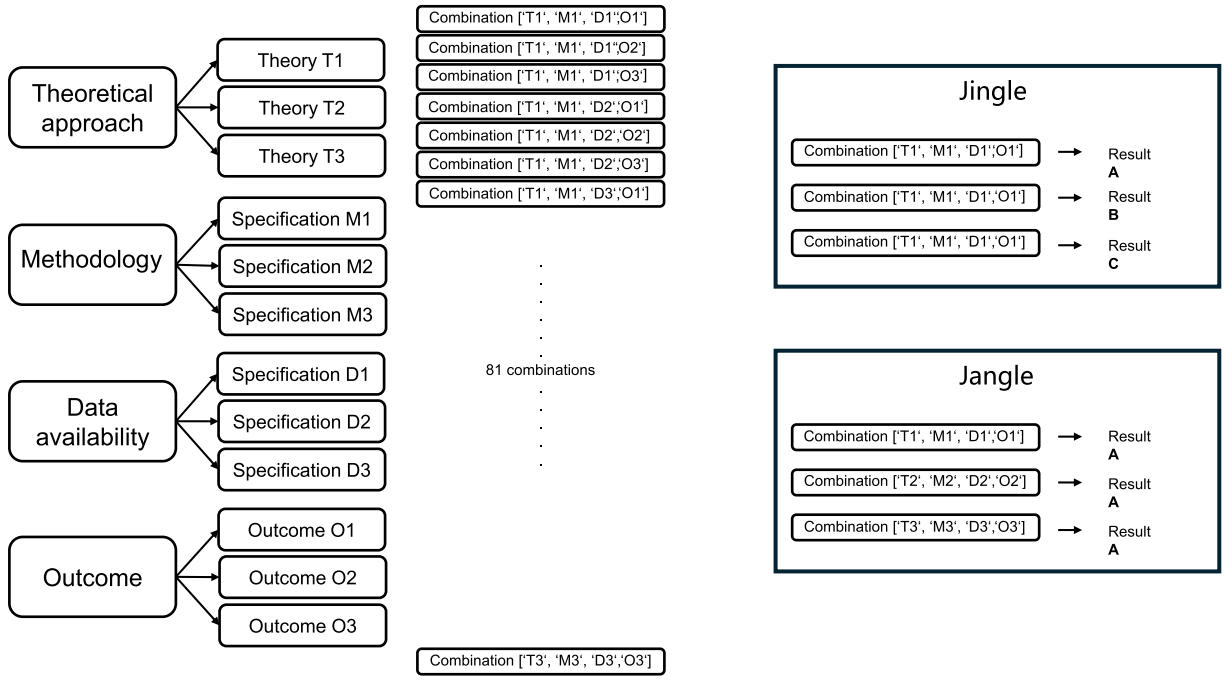



FIGURE 1
Exemplary specifications derived from theory, methodology, data availability, and results.

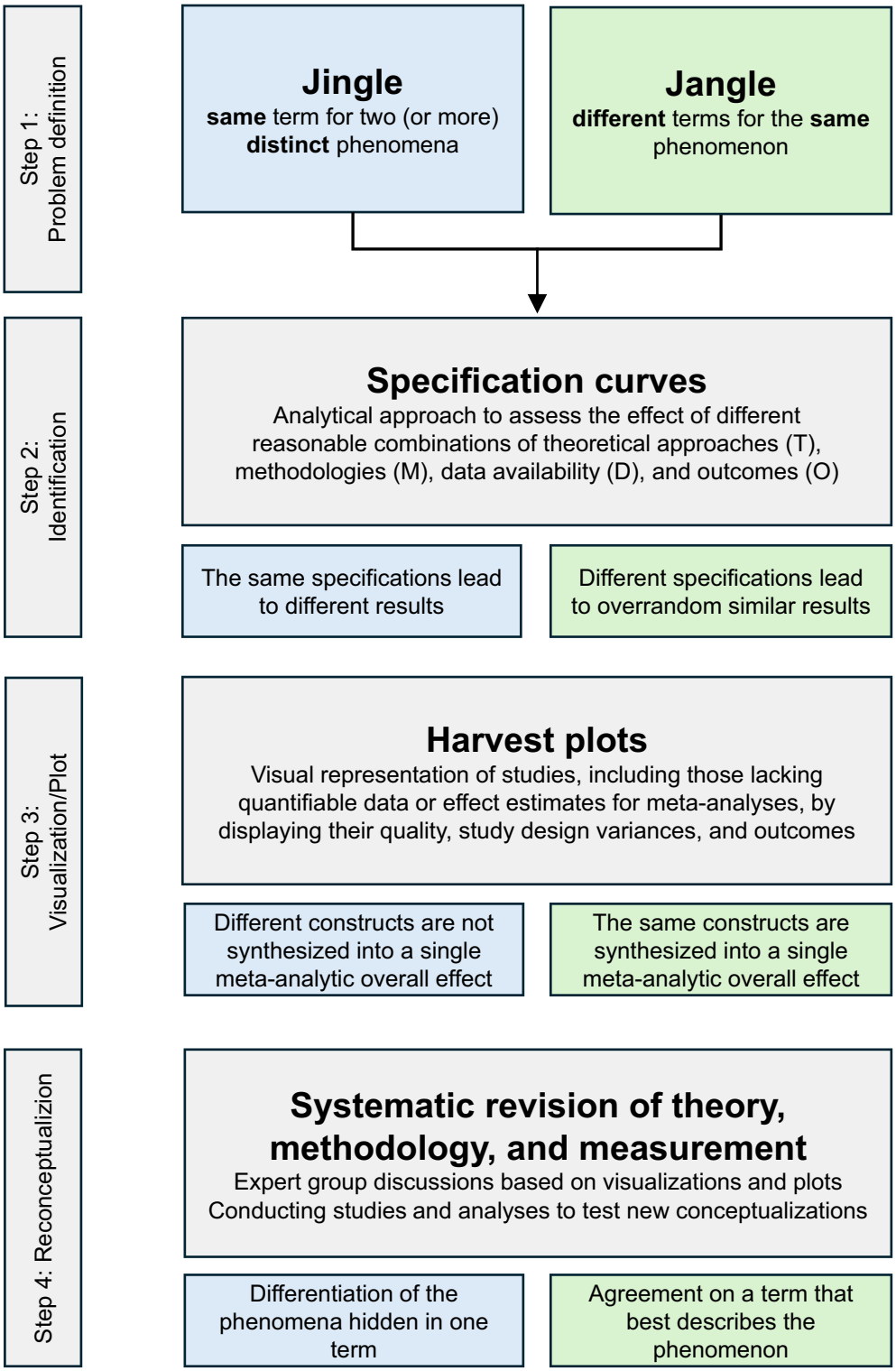



FIGURE 2
Process for systematically revising jingle and jangle fallacies.

provides, for example, specification curve analysis developed by Simonsohn et al. (2020). The procedure delineates all reasonable and debatable choices and specifications for addressing a research inquiry at the single-study level. These specifications must (1) logically examine the research question, (2) be expected to maintain statistical validity, and (3) avoid redundancy with other specifications in the array. Steegen et al. (2016) introduced the multiverse analysis concept, offering additional plotting alternatives as a similar approach. Voracek et al. (2019) combined these approaches at a meta-analytical level, revealing the range of formally valid specifications, including theoretical frameworks, methodological approaches, and researchers' degrees of freedom. They distinguish between internal ("which," e.g.,

the selection of data for meta-analysis) and external ("how," e.g., the methodology of data meta-analysis) factors. Identifying reasonable and formally valid specifications is considered a crucial first step in gaining an overview of which aspects and perspectives of a psychological phenomenon have already been empirically investigated.

Detecting and preventing jingle and jangle fallacies requires considering as many studies as possible to obtain a comprehensive overview. However, addressing the relatively strict and sometimes poorly justified inclusion and exclusion criteria in systematic reviews and meta-analyses presents a further challenge (Uttley et al., 2023). The current practice of setting rigid criteria in meta-analyses may be overly stringent, leading to the exclusion of valuable but non-quantifiable studies. Several approaches have less strict inclusion criteria, such as the harvest plot (Ogilvie et al., 2008). The harvest plot considers studies by graphical displays that otherwise would be excluded due to missing quantifiable data or effect estimates for meta-analyses, plotting the quality, the study design variances, differences of included variables, and outcome information of the studies. Foulds et al. (2022) described harvest plots as an exploratory method that allows for grouping outcomes, including non-parametric statistical tests, studies without effect sizes, and depiction of biases within studies. Comparing the results of a meta-analysis and a harvest plot analysis derived from the same study corpus reveals that the harvest plot approach allows for the inclusion of a significantly higher number of studies in the analysis (Foulds et al., 2022, Table 3). Accordingly, techniques like harvest plots play a vital role in expanding the scope of analyzed findings, which is crucial for achieving a comprehensive understanding of studies on a specific phenomenon.

## Implementing jingle and jangle detectors

As described, various useful approaches effectively structure and organize studies on a psychological phenomenon. But how can we detect potential jingle and jangle fallacies? Harvest plots summarize the findings of studies based on their suitability of study design, quality of execution, variance-explaining dimensions (such as gender and race), and outcomes quality (e.g., behavioral, self-reports). The plots offer descriptive representations and provide an overview of previously investigated results. Empty lines indicate missing data for known combinations of variables or specifications (see, e.g., Ogilvie et al., 2008, Figure 1). However, they still lack a combinatorial approach, as suggested by Simonsohn et al., 2020 or Voracek et al. (2019). After implementing the permutational aspect on specifications, a jingle fallacy detector could be based on the idea that, in the presence of a jingle, the same specifications would lead to different results. Conversely, jangle fallacy detectors would operate *vice-versa* and indicate jangle if different specifications yield overrandomly similar results. Figure 1 illustrates how combining different theoretical approaches, methodologies, data availabilities, and outcomes can help identify potential jingle and jangle fallacies.

Thus far, two promising approaches have concentrated on detecting jingle and jangle fallacies at a taxonomic level. Larsen and Bong (2016) presented six different so-called construct identity detectors for literature reviews and meta-analyses, applying different natural language processing algorithms. Wulff and Mata (2023) provided a solution in a preprint, utilizing GPT at the level of personality taxonomies to analyze the items and their scale assignments in the international personality item pool (IPIP; Goldberg et al., 2006). Since GPT is based on Natural Language Processing, it is well-suited to detect jingle and jangle fallacies within taxonomic approaches. However, reliance on taxonomies alone is insufficient for detecting jingle and jangle fallacies in psychological science. We understand psychological phenomena through theories, operationalized with concepts, constructs, and methodologies, and measured through physiological and behavioral data, self-reports, and external reports. Empirical data hinges on these interconnected elements alongside methodologies and study designs (Uher, 2023). Therefore, to detect jingle and jangle fallacies, we must consider all these levels and phases of psychological science.

## Conclusion

The growing attention to jingle and jangle fallacies in recent years underscores their significance in psychological science, posing a threat to validity and often going unrecognized. These fallacies, originally defined by Thorndike (1904) and Kelley (1927), emerge due to vague connections between theoretical concepts and empirical operationalizations but also have pure computational roots (Grieder and Steiner, 2022). Developments like meta-analyses and systematization through reviews help to systematize knowledge, but these practices are not immune to biases and limitations (Uttley et al., 2023) and do not detect jingle and jangle fallacies – such detectors are not yet developed. These detectors need to consider all levels and phases of psychological science, from theoretical frameworks to methodological approaches and study designs, called study specifications (Simonsohn et al., 2020). Additionally, flexible inclusion criteria for considered studies and new computational approaches, as conducted by Larsen and Bong (2016) or Wulff and Mata (2023) are needed. Ultimately, addressing jingle and jangle fallacies requires a concerted effort across the scientific community, incorporating diverse theories, perspectives, and methodologies. Simply defining the problem – finding one term for multiple phenomena (jingle) or different terms for the same phenomenon (jangle) – is insufficient. A systematic revision of jingle and jangle fallacies, achieved through discussion and analysis of detected instances is essential, as outlined in Figure 2.

## Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

## Author contributions

BH: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. SO: Writing – review & editing. JP: Writing – review & editing. UT: Writing – review & editing, Conceptualization. MV: Conceptualization, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note



## References

Altgassen, E., Geiger, M., and Wilhelm, O. (2024). Do you mind a closer look? A jingle-jangle fallacy perspective on mindfulness. *Eur. J. Personal.* 38, 365–387. doi: 10.1177/08902070231174575

Ayache, J., Dumas, G., Sumich, A., Kuss, D. J., Rhodes, D., and Heym, N. (2024). The «jingle-jangle fallacy» of empathy: delineating affective, cognitive and motor components of empathy from behavioral synchrony using a virtual agent. *Personal. Individ. Differ.* 219:112478. doi: 10.1016/j.paid.2023.112478

Beisly, A. H. (2023). The jingle-jangle of approaches to learning in prekindergarten: A construct with too many names. *Educ. Psychol. Rev.* 35:79. doi: 10.1007/s10648-023-09796-4

Elson, M. (2019). Examining psychological science through systematic meta-method analysis: a call for research. *Adv. Methods Pract. Psychol. Sci.* 2, 350–363. doi: 10.1177/2515245919863296

Fischer, A., Voracek, M., and Tran, U. S. (2023). Semantic and sentiment similarities contribute to construct overlaps between mindfulness, big five, emotion regulation, and mental health. *Personal. Individ. Differ.* 210:112241. doi: 10.1016/j.paid.2023.112241

Foulds, J., Knight, J., Young, J. T., Keen, C., and Newton-Howes, G. (2022). A novel graphical method for data presentation in alcohol systematic reviews: the interactive harvest plot. *Alcohol Alcohol.* 57, 16–25. doi: 10.1093/alcalc/agaa145

Glass, G. V. (2015). Meta-analysis at middle age: a personal history. *Res. Synth. Methods* 6, 221–231. doi: 10.1002/jrsm.1133

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *J. Res. Pers.* 40, 84–96. doi: 10.1016/j.jrp.2005.08.007

Gonzalez, O., MacKinnon, D. P., and Muniz, F. B. (2021). Extrinsic convergent validity evidence to prevent jingle and jangle fallacies. *Multivar. Behav. Res.* 56, 3–19. doi: 10.1080/00273171.2019.1707061

Grieder, S., and Steiner, M. D. (2022). Algorithmic jingle jungle: a comparison of implementations of principal axis factoring and promax rotation in R and SPSS. *Behav. Res. Methods* 54, 54–74. doi: 10.3758/s13428-021-01581-x

Hagger, M. S. (2014). Avoiding the "déjà-variable" phenomenon: social psychology needs more guides to constructs. *Front. Psychol.* 5:52. doi: 10.3389/fpsyg.2014.00052

Hanfstingl, B. (2019). Should we say goodbye to latent constructs to overcome replication crisis or should we take into account epistemological considerations? *Front. Psychol.* 10:1949. doi: 10.3389/fpsyg.2019.01949

Hook, J. N., Davis, D. E., van Tongeren, D. R., Hill, P. C., Worthington, E. L., Farrell, J. E., et al. (2015). Intellectual humility and forgiveness of religious leaders. *J. Posit. Psychol.* 10, 499–506. doi: 10.1080/17439760.2015.1004554

Kelley, T. L. (1927). Interpretation of educational measurements. New York and Chicago: World Book Company.

Lakens, D., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F. D., Hasselman, F., et al. (2017). *Examining the reproducibility of meta-analyses in psychology: A preliminary report*. Available at: https://doi.org/10.31222/osf.io/xfbjf

Larsen, K. R., and Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Q.* 40, 529–551. doi: 10.25300/MISQ/2016/40.3.01

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., and Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS One* 15:e0233107. doi: 10.1371/journal.pone.0233107

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., et al. (2019). The murky distinction between self-concept and self-efficacy: beware of lurking jingle-jangle fallacies. *J. Educ. Psychol.* 111, 331–353. doi: 10.1037/edu0000281

Ogilvie, D., Fayter, D., Petticrew, M., Sowden, A., Thomas, S., Whitehead, M., et al. (2008). The harvest plot: a method for synthesising evidence about the differential effects of interventions. *BMC Med. Res. Methodol.* 8:8. doi: 10.1186/1471-2288-8-8

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372:n160. doi: 10.1136/bmj.n160

Pigott, T. D., and Polanin, J. R. (2020). Methodological guidance paper: high-quality meta-analysis in a systematic review. *Rev. Educ. Res.* 90, 24–46. doi: 10.3102/0034654319877153

Polanin, J. R., Hennessy, E. A., and Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: a meta-review. *Perspect. Psychol. Sci.* 15, 1026–1041. doi: 10.1177/1745691620906416

Porter, T. (2023). Jingle-jangle fallacies in intellectual humility research. *J. Posit. Psychol.* 18, 221–223. doi: 10.1080/17439760.2022.2154698

Shadish, W. R., and Lecy, J. D. (2015). The meta-analytic big bang. *Res. Synth. Methods* 6, 246–264. doi: 10.1002/jrsm.1132

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nat. Hum. Behav.* 4, 1208–1214. doi: 10.1038/s41562-020-0912-z

Skinner, E. A. (1996). A guide to constructs of control. *J. Pers. Soc. Psychol.* 71, 549–570. doi: 10.1037/0022-3514.71.3.549

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* 11, 702–712. doi: 10.1177/1745691616658637

Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. New York: Teachers College, Columbia University.

Uher, J. (2023). What are constructs? Ontological nature, epistemological challenges, theoretical foundations and key sources of misunderstandings and confusions. *Psychol. Inq.* 34, 280–290. doi: 10.1080/1047840X.2023.2274384

Uttley, L., Quintana, D. S., Montgomery, P., Carroll, C., Page, M. J., Falzon, L., et al. (2023). The problems with systematic reviews: a living systematic review. *J. Clin. Epidemiol.* 156, 30–41. doi: 10.1016/j.jclinepi.2023.01.011

Voracek, M., Kossmeier, M., and Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Z. Psychol.* 227, 64–82. doi: 10.1027/2151-2604/a000357

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., and van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front. Psychol.* 7:1832. doi: 10.3389/fpsyg.2016.01832

Wulff, D. U., and Mata, R. (2023). Automated jingle–jangle detection: Using embeddings to tackle taxonomic incommensurability. *PsyArXiv*. doi: 10.31234/osf.io/9h7aw

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 02 September 2024
DOI 10.3389/fpsyg.2024.1437843



OPEN ACCESS

EDITED BY
Barbara Hanfstingl,
University of Klagenfurt, Austria

REVIEWED BY
Sara R. Jaeger,
Aarhus University, Denmark
Armand Cardello,
A.V. Cardello Editing and Consulting,
United States

*CORRESPONDENCE
Rainer Reisenzein
✉ rainer.reisenzein@uni-greifswald.de

RECEIVED 24 May 2024
ACCEPTED 30 July 2024
PUBLISHED 02 September 2024

CITATION
Reisenzein R and Junge M (2024) Measuring the intensity of emotions.
*Front. Psychol.* 15:1437843.
doi: 10.3389/fpsyg.2024.1437843



# Measuring the intensity of emotions

Rainer Reisenzein* and Martin Junge

Institute of Psychology, University of Greifswald, Greifswald, Germany

We describe a theoretical framework for the measurement of the intensity of emotional experiences and summarize findings of a series of studies that implemented this framework. Our approach is based on a realist view of quantities and combines the modern psychometric (i.e., latent-variable) view of measurement with a deductive order of inquiry for testing measurement axioms. At the core of the method are nonmetric probabilistic difference scaling methods, a class of indirect scaling methods based on ordinal judgments of intensity differences. Originally developed to scale sensations and preferences, these scaling methods are also well-suited for measuring emotion intensity, particularly in basic research. They are easy to perform and provide scale values of emotion intensity that are much more precise than the typically used, quality-intensity emotion rating scales. Furthermore, the scale values appear to fulfill central measurement-theoretical axioms necessary for interval-level measurement. Because of these properties, difference scaling methods allow precise tests of emotion theories on the individual subject level.

KEYWORDS

**emotion intensity, difference measurement, difference scaling, testing measurement axioms, indirect scaling methods, rating scales, emotion measurement**

## 1 Introduction

Linguistic and phenomenological evidence indicates that emotions—by which we mean emotional experiences—differ from each other not only in type or quality, but also in intensity. For example, we say not only that someone, including ourselves, is happy, sad, or surprised; we often qualify these emotion ascriptions with intensity modifiers such as "a little," "moderately," or "extremely": Karl is a little happy, Maria is moderately sad, we feel extremely surprised. These linguistic practices are supported by introspection, which confirms that different episodes of joy, sadness etc. can differ greatly in intensity, and that even during an emotion episode of constant quality, the intensity of the feeling can wax and wane. Generalizing these observations, one may say that linguistic and phenomenological evidence indicates that each emotion type can occur in different degrees or gradations, ranging from just noticeable to extremely intense.

This generalization suggests the hypothesis that emotions are *quantities*, that is, continuous magnitudes with an additive structure (see Michell, 1999 and Section 5). If so, theories of emotion should preferably be quantitative theories, that is theories in which magnitudes are connected by numerical functions (Carnap, 1966). However, stringent tests of these theories require measuring the intensity of emotions on a metric (interval or ratio) scale level. If emotions are indeed quantities, this should be possible in principle, i.e., provided suitable measurement methods can be devised. Indirect support for these assumptions is provided by the observation (Reisenzein, 2012) that, in being a group of related phenomenal qualities graded in intensity, emotions are similar to sensations (e.g., of tone, touch, or temperature).

Sensations, however, are generally regarded as quantities, and it is also widely believed that their intensity can be measured on a metric scale level (e.g., Stevens, 1975; Anderson, 1981; Schneider, 1982; Marks and Gescheider, 2002; Kingdom and Prins, 2010).

As a matter of fact, the assumption that emotional feelings, like sensations, are quantities whose intensity can therefore in principle be measured on a metric scale, has been made since the beginnings of academic psychology in the 19th century (e.g., Fechner, 1871; Külpe, 1893; Wundt, 1896; Titchener, 1902); and it continues to be held, at least implicitly, by probably most of today's emotion researchers. What is controversial, however, is how a precise, metric measurement of emotion intensity can be achieved.

This issue is particularly contentious regarding the most frequently used method for assessing the intensity of emotional feelings, the direct scaling of emotion intensity on quality-intensity rating scales (e.g., "How happy are you right now on a scale from 0 = not at all to 10 = extremely?"). The fact that most emotion researchers analyze these data with statistical methods that presuppose a metric scale level (e.g., linear regression), suggests that they believe that emotion intensity ratings are at least approximately metric. This view has been defended, for rating scales more generally, by several authors, most elaboratedly by Anderson (1981, 1982). In contrast, critics of rating scales insist, with equal tenacity, that rating scales are only ordinal and their analysis with metric statistical methods is therefore problematic, if not outright illegitimate (for a recent version of this critique see Liddell and Kruschke, 2018). Attempts to test the assumption that emotion rating scales—or, for that matter, other methods of measuring the intensity of emotional experiences—yield metric scales, are however exceedingly rare.

In view of the contested scale level of rating scales, as well as the many other criticisms raised against them (see Section 2.2.1), we have during the past years explored alternative methods of measuring emotion intensity that avoid the problems of rating scales and yield a metric scale level, or a least approach the metric level more closely than rating scales do (Junge and Reisenzein, 2013, 2015, 2016; Reisenzein and Franikowski, 2019; Reisenzein and Junge, 2024). As part of this research project, we also tested the metricity of emotion intensity ratings (Junge and Reisenzein, 2016). We found a suitable class of methods in probabilistic nonmetric difference scaling methods, a class of indirect scaling methods originally developed in psychophysics and preference measurement. Its main variants are *Ordinal Difference Scaling* (Agresti, 1992; Boschman, 2001; see also Tutz, 1986) and *Maximum Likelihood Difference Scaling* (Maloney and Yang, 2003; Knoblauch and Maloney, 2008). These methods have been successfully applied in the sensory and perceptual domain (e.g., Boschman, 2001; Maloney and Knoblauch, 2020), but prior to our studies, they were not used for the scaling of emotion intensity.

In this article, we summarize our research and elaborate and justify the theoretical approach to emotion intensity measurement that it exemplifies. Briefly, our approach is founded on a realist view of measurement (see, e.g., Michell, 1999, 2005; Tal, 2020) and combines the modern psychometric (i.e., latent-variable) approach to measurement (see, e.g., Borsboom, 2005) with a deductive order of inquiry of testing measurement axioms (Westermann, 1983, 1985). Although the components of this approach to mental measurement are not new, certain elaborations of these components are (see in particular Section 5.3), as is the application of the proposed method to the measurement of emotion intensity. The main part of the article describes our approach to emotion measurement and the findings obtained with it. This part is preceded by a brief review of scaling methods that have been used to measure the intensity of emotional experiences.

Before proceeding, it is important to emphasize that the proposed indirect scaling method is not intended to replace emotion ratings or other direct intensity scaling methods in all situations. As discussed in Section 5, difference scaling is not suitable for all measurement contexts, and is more costly than direct scaling methods. Nevertheless, we believe that in research contexts where difference scaling can be used, its additional costs are often a worthwhile trade-off for obtaining more precise, less biased, and closely metric measurements.

## 2 Methods for measuring the intensity of emotional experiences

### 2.1 The object of measurement: emotional states

When speaking of *emotions* in this article, we mean *occurrent emotional states*, such as an episode of joy, sadness, fear, or relief. Emotional states are temporary mental states of typically short duration, that are at least normally conscious, and are typically evoked by perceptions or thoughts of motivationally relevant objects or events. As conscious mental states, emotions are characterized by a more or less emotion-specific experiential quality that occurs with a particular intensity, and they are usually experienced as being directed at the evoking objects (e.g., Karl is happy about the arrival of a friend). Emotional states are what emotion psychologists are first and foremost interested in, and what theories of emotion are primarily about (see, e.g., Reisenzein, 2015). Our focus on emotional states means that we ignore here the measurement of *emotional dispositions*, i.e., tendencies or readinesses to have particular emotional states in suitable situations (see Reisenzein et al., 2020).

Although a definitive list of the mental states that count as emotions does not exist (see Reisenzein, 2012, for a discussion), there is broad agreement among emotion researchers, as well as lay people, on the core members of this list. These include joy and sadness, hope and fear, joy and pity for another, disappointment and relief, pride and anger, guilt, shame, disgust and many other mental states similar to these (see, e.g., Ortony et al., 1988). Because most of these mental states are subjectively experienced as either pleasant (e.g., joy, pride, relief) or unpleasant (e.g., sadness, fear, disappointment), having a definite hedonic tone has often – from Külpe (1893) to Ortony (2022) – been regarded as the decisive, or at least a central, criterion for being an emotion. The presence of a hedonic tone also justifies subsuming sensory pleasures and displeasures (the pleasant and unpleasant feelings evoked by colors, sounds, tastes, smells etc.) under the category of emotions, despite the fact that they differ from prototypical emotions in other respects (in particular, they have a less complex cognitive basis; see Ortony et al., 1988; Reisenzein, 2009). However, although having a definite hedonic tone may be sufficient for a mental state to qualify as an emotion, it is not universally regarded as necessary: Some theorists also regard certain mental states as emotions, or as emotion components, that do not appear to meet the hedonic criterion. Examples are surprise (e.g., Ekman, 1992; Reisenzein et al., 2019) as well as feelings of arousal (calm vs. aroused),

which several emotion theorists regard as a second basic feeling component of emotions (e.g., Wundt, 1896; Russell, 2003).

Disagreement about the classification of a mental state as an emotion is, however, not a hindrance to measuring its intensity, as long as it has an intensity at all. Nor is a worked-out scientific theory of an emotion needed to measure its intensity, at least for the self-report based measurement methods that are the focus of this article (Section 2.2): As long as the emotion whose intensity one wants to measure is known to common-sense psychology (Heider, 1958), it can be targeted in the measurement process by specifying it to competent language users with an appropriate, generally understood emotion term (e.g., "pleasure," "disgust," "fear," "relief").

## 2.2 Measuring emotion intensity by self-report: direct versus indirect scaling methods

In this article, we focus on psychometric scaling methods for the measurement of emotion intensity. These methods are ultimately based on introspection, and usually take the form of highly structured self-reports. To justify our focus on these methods, it would be sufficient to point out that they are the most frequently used methods for measuring the quality and intensity of emotions. But there are also important theoretical reasons for focusing on introspection-based methods of emotion measurement.

First, there is currently no objective indicator of emotions, whether physiological or behavioral, that can distinguish as finely between the different qualities and intensities of emotions as introspection-based self-reports can (see, e.g., Mauss and Robinson, 2009; Reisenzein et al., 2014). Second, and more fundamentally, introspective self-reports of emotion can claim epistemic priority over other emotion measurements. Even if one assumes that emotional states comprise more than just emotional experience, or that emotions can sometimes be unconscious (e.g., Plutchik, 1989), it is difficult to deny that the primary criterion for the presence, quality, and intensity of an emotion in a target person is the person's experience, to which the experiencer, and only the experiencer, has direct access. Indeed, it can be argued that the epistemic priority relation between introspective self-reports and other measures of emotions is inextricable: The science of emotion must accord epistemic priority to self-reports of emotional experience to maintain contact with the common-sense understanding of emotional states and their ascription (e.g., Heider, 1958; Laucken, 1974).

Although these arguments, particularly the second one, often evoke the dissent of emotion researchers when presented openly, they appear to be widely accepted implicitly. This is evidenced by the fact that self-report based measurements are typically used as the "gold standard" for validating behavioral and physiological measures of emotion (e.g., Reisenzein et al., 2014), and for selecting or constructing stimuli to induce emotions in laboratory studies (see also Kron, 2019).

The traditional aim of psychometric scaling methods has been to assess presumed mental quantities, such as sensations or emotional feelings, on a metric (interval or higher) scale level. In this article, we are only concerned with such attempts at metric measurement, i.e., measurement in the classical sense (Michell, 1999). The scaling methods that have been proposed for this purpose are often divided into "direct" and "indirect" methods (e.g., Engen, 1971; Sixtl, 1982; Marks and Gescheider, 2002). This distinction will also be used here because of its fundamental importance. The most important difference between direct and indirect scaling methods is how much of the process of constructing a metric scale is trusted to the subject.

### 2.2.1 Direct scaling methods for measuring emotion intensity

Direct scaling methods, when proposed for metric measurement, are based on the assumption that humans are in principle able to provide metric measurements of the intensity of their sensations and feelings, which can then be more or less directly used in subsequent data analyses. Direct scaling methods fall into two main classes, corresponding to the two main metric scale levels, interval and ratio, that their proponents believe can be attained with them (e.g., Engen, 1971).

The "intended interval-scale" methods, sometimes called *partition methods* (following Stevens, 1975), assume that people are able to partition the latent intensity continuum into a set of equal-sized intervals (e.g., Engen, 1971; Marks and Gescheider, 2002). The most prominent partition method is the category rating scale (Guilford, 1954). As mentioned, the quality-intensity emotion rating scale, or at least certain versions of this scale (Guilford, 1954; Anderson, 1981), are examples of this direct scaling method.

Despite their ubiquity and easy of use, rating scales in general have been extensively criticized (e.g., Baumgartner and Steenkamp, 2001; Marks and Gescheider, 2002; Yannakakis and Martínez, 2015; Uher, 2018, 2023), and these criticisms are also relevant for emotion rating scales (see also Lim, 2011). Probably the most important actual or potential problems of rating scales for the measurement of emotion intensity are (1) their limited resolution (see, e.g., Böckenholt, 2004); (2) their comparatively large contamination with random error, given typical and realistically possible conditions of use (see Section 3.4); (3) their nonmetric scale level, and, partly responsible for it (4) their susceptibility to stimulus and instructional context effects, as well as to diverse response biases (e.g., Poulton, 1989; Baumgartner and Steenkamp, 2001), such as the tendency to avoid the extremes of the scale (Stevens and Galanter, 1957). It should be noted, however, that there are ways to reduce context effects and response biases (Anderson, 1982) and that the influence of some commonly claimed rating response styles, such as acquiescence (Baumgartner and Steenkamp, 2001), on emotion intensity ratings appears to be minimal in typical assessment contexts (see Schimmack et al., 2002).

The "intended ratio-scale" class of direct scaling methods comprises various forms of *magnitude scaling*, which gained prominence primarily because of S. S. Stevens' psychophysical research (e.g., Stevens and Galanter, 1957; Stevens, 1975). The most frequently used magnitude scaling method is *magnitude estimation*, where participants are required to judge the ratio of the intensity of a sensation or feeling to an experimenter-supplied or (implicitly) self-chosen comparison standard.

Magnitude scaling methods have become highly popular in the field of sensory measurement (e.g., the measurement of sound intensity or brightness; Marks and Gescheider, 2002), not last because they were advertised as superior to category ratings (Stevens and Galanter, 1957; Stevens, 1975). Nonetheless, magnitude scaling methods have only been rarely used for the measurement of emotion intensity (examples are Moskowitz and Sidel, 1971; Sullivan, 1971; Teghtsoonian and Frost, 1982; Galanter, 1990; see also Cardello and

Jaeger, 2010; Lim, 2011). The main reason for this neglect may have been practical: As Cardello and Jaeger (2010) and Lim (2011) point out for the field of sensory science, some participants have difficulties learning magnitude estimation procedures, and the resulting data are more cumbersome to process than ratings. In addition, the claimed advantages of magnitude scaling over category ratings—that magnitude scaling is immune to stimulus and instructional context effects, and yields a ratio scale (Stevens, 1975)—have turned out to be highly questionable (see Anderson, 1981; Birnbaum, 1982; Ellermeier and Faulhammer, 2000; Masin, 2022).

### 2.2.2 Indirect scaling methods for measuring emotion intensity

The criticisms of the direct scaling methods are good reasons to consider *indirect scaling methods* as alternative methods for measuring emotion intensity. Two common arguments for indirect scaling methods are that they are less susceptible to response biases (see, e.g., Brown and Maydeu-Olivares, 2018) and that they yield more precise measurements than direct scalings. Historically, however, the central motivation for developing indirect scaling methods was the belief that direct scaling methods cannot provide metric measurements, whereas indirect scaling methods can.

The most conservative indirect scaling position is that people's introspective abilities are limited to judging the intensities of the sensations or feelings evoked by different stimuli on an ordinal scale, i.e., as greater, equal or less (e.g., Fechner, 1860, 1871; Thurstone, 1927). A more optimistic view, apparently first articulated by Plateau (1872), is that people can additionally order *differences* between feeling intensities. We come back to this assumption in Section 3.5. The important point at present is that, in both cases, the introspecting subject is assumed to be only able to operate on the ordinal level of measurement: to rank-order intensities, or to (also) rank-order intensity differences. It is the researcher, who—on the assumption that the ordinal judgments are based on a latent quantitative variable—attempts to infer the exact levels of this variable from the ordinal judgments. This is achieved by using a scaling model (e.g., Thurstone, 1927; Boschman, 2001; Marley and Louviere, 2005). Interpreted from the realist view of mental measurement (see Sections 3.6 and 5.2), a scaling model is a theory about how (by which cognitive processes) the person's overt judgments are constructed on the basis of—in the case of conscious mental states—her introspective observation of the latent quantity. The process of estimating scale values attempts to invert the hypothesized judgment process, i.e., to estimate the values of the latent variable from the ordinal data plus the scaling theory's assumptions about the judgment process. An example of an indirect scaling model, the ODS model, is described in Section 3.6.

As said, psychometric emotion measurement is today dominated by a direct scaling method, the quality-intensity emotion rating scale. But this was not always so. To the contrary, at the beginnings of psychology as an academic discipline in the 19th century, indirect methods of measuring sensations and feelings predominated. The reason was that most psychologists of this period, despite regarding introspection as psychology's main method, did not believe that the intensity of sensations and feelings can be *directly* scaled. The first application of an indirect scaling method to emotional experience was made by Fechner (1871), who proposed an early version of best-worst scaling (see Cardello and Jaeger, 2010) to measure the aesthetic pleasantness of geometric figures. Somewhat later, Cohn (1894) used the paired comparison method (see Thurstone, 1927) to measure the pleasantness of colors, and Titchener (1902) extended the method to the measurement of the basic feelings postulated in Wundt's (1896) tri-dimensional theory of emotions. These early applications of indirect scaling methods to emotions were not based on an explicit scaling model; instead, scale values were estimated using intuitively plausible, simple calculations, such as counting how often each stimulus is judged as more pleasant than others. It was left to Thurstone (1927) to supply one of these methods, the paired comparison procedure, with an explicit statistical judgment (scaling) model that promised to yield metric measurements, provided that its assumptions are met. Thurstone's (1927) publication led to a rapid increase in the use of the paired comparison method for measuring attitudes, values and hedonic feelings (Cardello and Jaeger, 2010).

Although direct scaling methods for measuring emotional feelings, in the form of the category rating scale, also have a long history (Major, 1895), they gained popularity only in the 1940ies and 1950ies (Lorr, 1989; Cardello and Jaeger, 2010). They were first utilized more widely in the field of sensory hedonics, where the so-called *9-point hedonic scale*, a bipolar labeled category scale ranging from "dislike extremely" to "like extremely" (Peryam and Pilgrim, 1957) became dominant (Cardello and Jaeger, 2010). The main reason for its rise in popularity was practical: For measuring people's hedonic reactions to foods, beverages etc., the paired comparison method was experienced as too cumbersome or even inapplicable (Cardello and Jaeger, 2010).

It was only in the mid-1950ies to early 1960ies that researchers became interested in the assessment of specific emotions and moods (e.g., Nowlis and Nowlis, 1956; see Lorr, 1989). When they did, they turned to the quality-intensity rating scale almost by default. The main reason was again most likely practical: Rating scales are well-suited for the quick and comprehensive assessment of a person's momentary emotions or moods, which was then a major research interest (Nowlis and Nowlis, 1956). Still, it is worth noting that indirect scaling methods were not even considered anymore when emotion researchers began to measure specific emotions. From the beginning, nearly all attempts to measure specific emotions have used direct scaling methods—essentially some version of the ubiquitous quality-intensity rating scale.

### 2.2.3 More recent developments

Over the past two decades, the firm grip of the classical rating scale on emotion measurement has begun to loosen a little, due to the emergence of several new or improved direct and indirect scaling methods. Perhaps the most noteworthy development in the direct scaling camp is a new type of labeled intensity rating scale, where the placement of the intensity labels is determined empirically through magnitude estimation. These scales are known as *labeled affective magnitude scales* (e.g., Schutz and Cardello, 2001; Lishner et al., 2008; for reviews, see Cardello and Jaeger, 2010; Lim, 2011; Schifferstein, 2012; Ares and Vidal, 2020). Although it seems that these scales have so far only been used to measure the intensity of pleasure and displeasure, they could easily be adapted to assess specific emotions.

In the indirect scaling camp, too, new methods have been proposed to measure emotion intensity. Particularly noteworthy is *Best-Worst Scaling*, a modern probabilistic version of the scaling procedure proposed by Fechner (1871) (Finn and Louviere, 1992; for more recent accounts, see Marley and Louviere, 2005; Jaeger et al.,

2008; Louviere et al., 2015). This scaling method has become increasingly popular during the past years for measuring preferences and attitudes in several disciplines (see Schuster et al., 2024) and has also been utilized to measure emotions. So far, the focus of Best-Worst scaling in this area has been the measurement of sensory pleasure and displeasure (e.g., Jaeger et al., 2008; Jaeger and Cardello, 2009; Mielby et al., 2012; see also Cardello and Jaeger, 2010), but it has also been used to measure the intensity of fear (Farkas et al., 2021) and to scale the intensity of positive and negative emotions expressed in text (Mohammad and Bravo-Marquez, 2017). We believe that the probabilistic difference scaling methods advocated in this article (see Section 3) represent an even more effective indirect scaling alternative for measuring the intensity of emotions.

# 3 Difference scaling methods

## 3.1 Difference data

Difference scaling methods are indirect, unidimensional scaling methods based on difference data. Difference data (in our case, judgments) come in two main kinds: direct difference comparisons or *quadruple judgments* (QCs), and *graded paired comparisons* (GPCs). Both judgment tasks are special forms of the paired comparison method. In the QC task—the classical difference judgment task—the participants are in each trial presented with two pairs of stimuli (*a*, *b*) and (*c*, *d*) and indicate which pair differs more on the judgment dimension. For example (Junge and Reisenzein, 2015, Study 1), participants are shown two pairs of disgusting pictures side by side on the screen, and are asked to indicate in which pair the stimuli differ more in the intensity of evoked disgust.

In contrast, in the GPC task, two stimuli *a* and *b* are compared, as in the classical paired comparison task (e.g., Cohn, 1894; Thurstone, 1927). However, different from classical paired comparisons, the participants indicate not only which stimulus has the larger value on the judgment dimension, but also how much greater the difference is. Importantly, nonmetric scaling methods for GPCs assume that these judgments have only an ordinal scale level. So understood, the GPC task can be seen as a combination of the classical paired comparison task with an ordinal rating of differences. To illustrate, in another part of their Study 1, Junge and Reisenzein (2015) presented participants with the disgusting pictures in pairs and asked them to indicate which picture was more disgusting, as well as how much more disgusting it was, on a response scale with six ordered categories ranging from "just barely noticeably more" to "extremely more."

## 3.2 Scaling models for difference data

For both QCs and GPCs, a number of scaling methods are available. Here, we only consider nonmetric methods. In the first empirical studies using difference scaling, unidimensional versions of nonmetric multidimensional scaling were used (see, e.g., Schneider, 1982 for QCs and Orth, 1982, for GPCs). A disadvantage of these methods is, however, that they are not based on a statistical model (for additional discussion, see Haghiri et al., 2020). This drawback has been rectified in more recent, probabilistic scaling models whose main varieties are *Maximum Likelihood Difference Scaling* (MLDS) for QCs (Maloney and Yang, 2003; Knoblauch and Maloney, 2008), and *Ordinal Difference Scaling* (ODS) for GPCs (Agresti, 1992; Boschman, 2001; see also Tutz, 1986). These two scaling methods are actually closely related in terms of their basic assumptions (see Junge and Reisenzein, 2015). Furthermore, because both methods were developed for the scaling of ordinal difference data, both can claim to be founded on an axiomatic measurement theory developed for such data, the difference measurement model (Krantz et al., 1971). This means that ODS and MLDS not only allow to estimate precise scale values and to determine the overall fit of the model to the data, but also to construct a statistical test of the crucial axioms of difference structures that need to be fulfilled to obtain a metric scale (see Section 5).

## 3.3 Advantages of ODS over MDS

Although both MLDS and ODS are suitable for the measurement of emotion intensity (Junge and Reisenzein, 2015), in our studies we focused on ODS of GPCs, because this method has several advantages over MLDS, particularly for emotion measurement (Junge and Reisenzein, 2015). Most importantly, ODS is more economical than MLDS, because it needs much fewer input data (for details, see Junge and Reisenzein, 2015; and Schneider, 1982). This is a direct consequence of the fact that the input data of ODS (i.e., GPCs) require comparing pairs of stimuli, whereas those of MLDS require comparing pairs of pairs. The savings in the number of paired comparisons enabled by GPCs are substantial and increase with the number of stimuli (see Junge and Reisenzein, 2015). Additionally, because the GPC task requires processing only two stimuli rather than four in each trial, as the QC task does, it is arguably less cognitively taxing for the participants (Junge and Reisenzein, 2015). Finally, MLDS in contrast to ODS requires that the rank-order of the stimulus intensities is known, which in the case of affective stimuli usually means that this rank order has to be separately estimated for each participant prior to the QC task.

Importantly, the economical advantage of ODS does not come at the expense of lower-quality scalings: Junge and Reisenzein (2015) found that ODS scalings of GPCs were at least as reliable, and correlated at least as highly with direct ratings of emotion intensity, as MLDS scalings of QCs of the same stimuli. Hence, ODS can be regarded as an economical alternative to MLDS for the difference scaling of emotion intensity.

## 3.4 Differences to classical Thurstonian scaling

Although ODS and MLDS stand in the tradition of Thurstonian scaling models (Thurstone, 1927; Böckenholt, 2006), they differ in a crucial respect from other models of this class, including best-worst scaling (Marley and Louviere, 2005): They use not only information about the ordering of stimulus intensities, but also about the ordering of intensity differences. This additional information leads to several advantages of difference scaling methods (see Knoblauch and Maloney, 2008, for the case of MLDS; and also Anderson, 1981) that we here illustrate by comparing them to Thurstone's (1927) classical paired comparison model. First, in contrast to the Thurstonian model,

difference scaling models allow to test measurement axioms required for a metric representation (Krantz et al., 1971; see Section 5). Second, they allow to scale stimuli with clear suprathreshold intensity differences, i.e., stimuli that are perfectly discriminable, whereas the Thurstonian model can only estimate distances between stimulus pairs that are close enough to be not consistently distinguishable. Third, the difference scaling models allow to scale the data of individual participants, because a single judgment of the stimulus pairs or quadruples is sufficient to obtain reliable scale estimates. In contrast, Thurstonian scaling of individual data is unfeasible for many kinds of stimuli, because it requires numerous repetitions of the paired comparisons to obtain reliable estimates of the confusion probabilities (Anderson, 1981). Fourth, in the Thurstone model, the obtained scale depends crucially on the assumed error distribution, whereas MLDS has been found to be robust to variations of the error distribution (Maloney and Yang, 2003), and we have found the same for ODS in additional analyses of our data. Finally, whereas the interpretation of intervals on the MLDS and ODS scales as intensity differences is transparent, an analogous interpretation of the intervals on the confusion-based Thurstone scale requires additional assumptions (Knoblauch and Maloney, 2008).

## 3.5 Are people able to order intensity differences?

The information that difference scaling methods attempt to elicit from participants was first described by Plateau (1872) in a seminal paper on the measurement of sensations. In this article, Plateau (1872) conjectured:

> "When we experience, either simultaneously or successively, two physical sensations of the same sort, but of different intensities, we can easily judge which of the two is the stronger and, we can, moreover, decide whether the difference between them is great or small. But there, it seems, the comparison must end…we appear to be incapable of estimating the numerical ratio between the two intensities of two sensations in this way" [Plateau, 1872, translation by Laming and Laming (1996); p.136]

Note that the GPC task nearly precisely matches Plateau's (1872) description of what humans are, in his view, able to provide: Information about the ordering of the intensity of the compared sensations or feelings, and information about the ordering of their intensity differences ("barely different," "moderately different," "very different" etc.). Note also that Plateau's (1872) views on people's judgment abilities provide a precise explanation of the intuition behind the commonly made claim that rating scales are somewhere between the ordinal and metric scale levels, i.e., that they contain more than ordinal information, even though not metric information: People are also able to order the intensity differences between different sensations or feelings.

Are Plateau's assumptions plausible, and hence, can difference scaling work in principle? His first assumption, that people can reliably rank-order the intensity of the sensations or feelings evoked by different stimuli, is largely uncontroversial, provided that the intensity differences are not too small. However, for GPCs, this assumption can also be checked by testing the transitivity of the dichotomized GPC judgments. For the GPCs of emotion intensity collected in our studies, this analysis (conducted for the present article) revealed that the judgments were nearly perfectly transitive for practically all participants.

Thus, the validity of the GPC (and, analogously, the QC) method depends on Plateau's (1872) second assumption, that people are also able to consistently order intensity differences. As discussed in Section 5, this is still not enough; the ordering of intensity differences must also fulfill an additivity condition. However, already the more basic ability to order intensity differences has been questioned by some authors. Specifically, in the field of preference measurement, where axiomatic difference measurement has been a major research topic (for reviews, see, e.g., Krantz et al., 1971; Farquhar and Keller, 1989; Köbberling, 2006; Moscati, 2019), some researchers have doubted that people are able to compare and order preference differences (e.g., Machina, 1981). However, other researchers in this field have argued that this doubt is unfounded, that people are well able to order preference differences, and that the obtained data make sense (von Winterfeldt and Edwards, 1986). In any case, there is empirical evidence that people are able to provide reliable judgments of intensity differences of *sensations* and *emotional feelings* (e.g., Schneider, 1982; Knoblauch and Maloney, 2008; Junge and Reisenzein, 2015).

While these data are ultimately decisive, to convince oneself that people are indeed able to order the size of emotion intensity differences, it is best to consider an example (see also Krantz et al., 1971, p. 140–141, and von Winterfeldt and Edwards, 1986, pp. 209–210, who discuss similar examples). Imagine you are shown three affective pictures *a*, *b*, *c*, and find that they evoke, in order, just noticeable pleasure (say 1 on a 0–10 rating scale), mild pleasure (3), and very strong pleasure (9). As mentioned in the introduction, such intensity judgments of emotion are commonly made in everyday life, although not usually on a rating scale. Then ask yourself whether you would be willing to say that the difference between *b* and *c* (between mild and very strong pleasure) is greater than that between *a* and *b* (just noticeable and mild). If you answer yes (as we do), you agree that intensity differences of pleasure can be rank-ordered.

## 3.6 ODS as a psychological measurement theory

### 3.6.1 The ODS model

On a realist interpretation of measurement (see Section 5), the statistical model underlying ODS is a small psychological theory of the mental processes that underlie responses in the GPC task. (The same is true for the MLDS model of the QC task; see Junge and Reisenzein, 2015). The ODS model can be summarized in two equations:

$$\Delta_{a,b} = \Psi_b - \Psi_a + \varepsilon, \text{with } \varepsilon \sim N\left(0,\ \sigma^2\right) \tag{1}$$

$$R_{a,b} = j \text{ if } \theta_{j-1} < \Delta_{a,b} \leq \theta_j, \text{with } j = 1,\ldots,J \text{ and } -\infty = \theta_0 < \theta_1 < \ldots < \theta_{J-1} < \theta_J = +\infty \tag{2}$$

$\Psi_a$ and $\Psi_b$ are the scales values of the two stimuli *a* and *b* compared in a trial of the GPC task, and $\Delta_{a,b}$ is an internal decision variable on which the overt response $R_{a,b}$ is based. In addition, the ODS model contains $\theta_1$, …, $\theta_{J-1}$ unknown thresholds separating the response categories, which, like the scale values, must be estimated.

Interpreted in terms of mental processes, and illustrated for emotion intensities, the ODS model can be described as follows. Equation 1 describes the initial stimulus representation and comparison process. It assumes: (1) the emotion intensities evoked by the two stimuli *a* and *b* presented to the participant in a trial of the GPC task give rise to two emotion intensities whose values are on average $\Psi_b$ and $\Psi_a$. (2) The emotion intensities are compared, either simultaneously or successively, by a process that (implicitly) computes the difference between them (see 3.6.2 for an explication of this process). (3) Both processes (the elicitation of the feelings and their comparison) are biased by independent random noise stemming from a normal distribution with constant variance $\sigma^2$. Note, however, that the distributional assumption can be changed, and the constant variance assumption can in principle be relaxed.

Equation 2 describes the response process. It assumes: (4) The decision variable $\Delta_{a,b}$, which represents the computed difference between the intensities of the emotions elicited by stimuli *a* and *b* in a given trial, is mapped into category j of the response scale consisting of J ordered categories, whenever $\Delta_{a,b}$ lies between the thresholds $\theta_{j-1}$ and $\theta_j$ that mark the boundaries of j on the latent continuum. If the judgment noise were zero, the difference between the two intensities would be exactly mapped into the correct response category; however, because of the presence of random noise, another response category will occasionally be chosen, and this will happen more frequently, the closer the intensities evoked by the two stimuli are on the judgment dimension.

The aim of ODS scaling is to estimate, from the observable responses $R_{ab}$ (the ordinal graded comparisons of stimuli *a* and *b*), the latent scale values of the stimuli assumed to underlie these responses.

As just described, the ODS model is a special case of the ordered (or cumulative) probit model (McKelvey and Zavoina, 1975; Greene and Hensher, 2010), which can be obtained in a straightforward manner by applying the ordered probit model to GPCs (Agresti, 1992; Boschman, 2001; as pointed out by Agresti (2010), the proportional odds assumption characteristic for cumulative link models is implied by a simple latent variable model). The scale values and thresholds can be estimated using maximum likelihood methods with widely available software. For example, in R (R Core Team, 2023), one can estimate the ODS model parameters with the functions *polr* in library *MASS* and *clm* in library *ordinal* (Christensen, 2018). Functions for the Bayesian estimation of the ordered probit model are also available (e.g., Gelman and Hill, 2006; Bürkner, 2017). In our research with ODS, we estimated the ordered probit model using a bias-reducing version of maximum likelihood estimation, *bpolr* (Kosmidis, 2014). This was done to avoid issues of separation, an estimation problem that can occur particularly with sparse data, e.g., when estimating the model for individual subjects (for more information, see Junge and Reisenzein, 2015).

### 3.6.2 Possible elaborations of the ODS model

As it stands, the ODS model is a relatively coarse and abstract theory of the mental processes that take place in the GPC task. Elaborations of the model are possible, however, two of which we sketch here.

First, one could refine the ODS model by distinguishing between assumed subprocesses. In particular, one could introduce a threshold for noticing intensity differences, and one could try to tease apart the different sources of random noise that contribute to the error term and model them by separate parameters. These noise sources are in particular (a) trial-by-trial fluctuations of the emotion intensities evoked by a stimulus (e.g., because of different degrees of attention devoted to the stimulus in different trials); (b) fluctuations due to the limited precision of the difference comparison mechanism; (c) fluctuations in the mapping of the decision variable to the response categories; and (d) response errors due to lapses of attention or wrong key presses. This general path to model elaboration has been taken in other areas of psychometric modeling, for example in models for temporal order and simultaneity judgments (e.g., García-Pérez and Alcalá-Quintana, 2012; see also Reisenzein and Franikowski, 2022). Its practical advantage for measurement is that, by isolating the different component processes and estimating them separately, purer estimates of the latent emotion intensities can be obtained.

Second, one could elaborate the ODS model into a full-fledged cognitive process model, that is, a representational-computational model of the judgment process. This requires specifying the underlying representation medium or media and the basic operations performed with these representations during the judgment process. A computational model does not at present exist for GPC (nor QC) judgments. However, Petrov and Anderson (2005) have proposed a computational model for category ratings in the well-researched ACT-R cognitive architecture (e.g., Anderson, 1983; Anderson and Lebiere, 1998). This computational model, which combines the Thurstonian theory of category ratings (Torgerson, 1958) with the theory of memory incorporated in the ACT–R architecture, could serve as the template for an analogous computational model of the GPC task. We briefly sketch here how this model might look like, because doing so adds substance and plausibility to our realistic interpretation of ODS as a psychological judgment theory.

Following analogous assumptions by Petrov and Anderson (2005) for the category rating task, we begin by assuming that the first step of the GPC task is the creation of emotion intensities for the two stimuli *a* and *b* compared in a trial. The details of this process need not be specified for measurement purposes, with one exception: We assume that these intensities are a form of analog representation of magnitudes (see Beck, 2015, for more on this concept). The two intensity representations are then processed within the central subsystem of ACT-R. The first central processing step is the computation of the intensity differences. We propose that this is achieved by a subpersonal similarity matching process, as implemented in the ACT-R architecture; hence it does not require symbolic (propositional) representations. Because the two intensities lie on an unidimensional quality continuum, the similarity comparison process amounts to a comparison of the intensities of the emotions (see already Thurstone, 1927). Furthermore, we submit that the resulting difference representation is again nonpropositional: It is an analogical representation of perceptual closeness or distance subjectively experienced as a feeling of smaller or greater difference.

This difference representation is next compared by the partial matching mechanism to a set of memory anchors that encode prototypical degrees of intensity differences more or less specific to the

emotion in question (see Petrov and Anderson, 2005). More precisely, the difference representation activates an anchor whose magnitude is similar to the computed intensity difference. Anchor selection is stochastic and also depends on other factors besides similarity, such as recency and base-level strength. Furthermore, following once more Petrov and Anderson (2005), we may assume that, if there is a large discrepancy between the difference representation and the magnitude of the anchor retrieved from memory, an explicit correction mechanism may increment or decrement the response suggested by the anchor. Finally, one could include a learning mechanism that causes slight changes of the magnitude of the anchor that corresponds to the response in this trial (Petrov and Anderson, 2005).

## 3.7 Estimating the zero point

Unless special measures are taken, ODS—like all comparative judgment methods (Guilford, 1954; Böckenholt, 2004)—does not estimate the zero point of the scale. However, for many research questions of emotion psychology, it is at least advantageous, if not necessary, to also know the natural zero point (the absence of emotion), and thus to have available not just an interval scale (see Section 5) but a ratio scale. For example, a ratio scale of emotion intensity is needed for stringent tests of quantitative emotion models (e.g., Junge and Reisenzein, 2013, Study 1).

In our studies, we estimated the zero point using simultaneously collected direct ratings of emotion intensity. These ratings were made on numerical scales anchored at the lower end by the natural zero point of emotion intensity (e.g., "the picture evokes no pleasure") and at the upper end by "extremely intense." To locate the zero point on the ODS scale, we then transformed the ODS scale values into the range of each participant's ratings. Note that this method of estimating the zero point only relies on the ratings for estimating the distance from zero of the lowest-intensity stimulus. The error of this estimate will be minimal if that stimulus is indeed close to zero (i.e., if a low-intensity stimulus is in the set), which was almost always the case in our studies. However, it is also possible to estimate the zero point as part of the difference scaling procedure. The simplest way to achieve this is by including an affectively neutral (at least with respect to the emotion under study) stimulus, such as an affectively neutral picture. Additional methods for estimating the zero point of scales derived from comparative judgments are discussed by Guilford (1954) and Böckenholt (2004).

While the natural zero point of emotion intensity is the same for different people, to optimize the interpersonal comparability of emotion intensity scales, it would be ideal to also have an interpersonally comparable scale unit. For some research questions, this is even necessary (see, e.g., Bartoshuk et al., 2005; Luce, 2010; Schifferstein, 2012). A fully satisfactory solution to this problem does not exist. However, a pragmatic solution is to fix the scale unit by using an approximately consensual end-point anchor label, such as "maximal" or "extremely," on a parallel rating scale (see Borg, 1962; Marks et al., 1983). This approach is, in fact, common practice for labeling emotion rating scales. Sometimes, in particular when using imagined emotion-evoking scenarios, it is also possible to include a stimulus into the difference scaling procedure that can be assumed to evoke near-maximum emotion intensity in most people (Reisenzein and Junge, 2024). Another possibility may be to fuse difference scalings with data from cross-modality matching (Bartoshuk, 2014).

# 4 Measuring emotion intensity with difference scaling methods

In our studies, participants made GPC judgments of the intensities of a broad range of emotions: pleasure and disgust evoked by affective pictures, amusement and surprise induced by quiz items, relief and disappointment about lottery outcomes, hope and fear, disappointment and relief experienced in diverse imagined scenarios, and anger and pity in hypothetical helping situations (Junge and Reisenzein, 2013, 2015, 2016; Reisenzein and Franikowski, 2019; Reisenzein and Junge, 2024). In all studies, the participants also made direct scalings of emotion intensity on 0–10 or 0–100 numerical rating scales ranging from "not at all" to "extremely"; in one case, a combination of rating and ranking (Kim and O'Mahony, 1998) was used. In the studies reported in Junge and Reisenzein (2016), we additionally collected QC judgments, i.e., direct comparisons of intensity differences.

The GPC judgments were scaled with ODS and/or, in some cases, with MLDS, taking advantage of the fact that GPCs can be expanded to QCs, the data needed for MLDS (Junge and Reisenzein, 2013, 2015; see Section 5). The difference scaling models were fitted to the data of the individual participants and the estimated scale values were linearly transformed into the range of the rating scale to estimate the zero and an extreme point, and thus, improve the interpersonal comparability of the measurements.

## 4.1 Reliabilities and discrimination capacity

Across the studies conduced by Junge and Reisenzein (2013, 2015, 2016), the difference scalings of the individual participants had an average reliability (estimated either by repeated measurements, or a bootstrap procedure) of $r = 0.95$. In contrast, the average reliability of the ratings (estimated as the re-test correlation between ratings made before and after the GPCs, or in two different sessions) was $r = 0.79$. Furthermore, whereas the 0–10 category rating scale used in most of our studies allowed the participants to distinguish, at best, between one scale point, additional analyses revealed that the difference scale (transformed into the same range) enabled them to reliably distinguish between about 0.5 scale points.

In unpublished research, similar findings were obtained for ODS scalings of hope, disappointment, fear and relief in hypothetical scenarios (Reisenzein and Junge, 2024) and for feelings of pity and anger toward others in helping scenarios (Reisenzein and Franikowski, 2019).

## 4.2 Robustness of GPC scalings to variations of the difference scaling method

Scalings of the GPCs by ODS and by MLDS (after expanding the GPCs to QCs; see Section 5) yielded nearly identical results, with average intra-individual scale intercorrelations of $r > 0.99$ (Junge and Reisenzein, 2015). Additional analyses conducted by us on the data from Junge and Reisenzein (2015) found equally high correlations

between the ODS scale values and those estimated by a metric version of difference scaling, additive functional measurement (AFM, Boschman, 2001). This replicates findings by Boschman (2001) obtained for the scaling of sensory attributes. Junge and Reisenzein (2013) obtained slightly lower (average intra-individual $r = 0.95$) correlations between MLDS and AFM scalings. Taken together, these findings support the robustness of the GPC scaling results to variations of the probabilistic difference scaling method.

## 4.3 Testing emotion theories with difference scalings

Junge and Reisenzein (2013) used the MLDS and AFM models as auxiliary measurement theories to test two small psychological emotion theories. The intensities of the emotions were first estimated using difference scaling on the individual level, and these measurements were then used in experimental tests of the emotion theories. This sequential approach (measurement—theory test) corresponds to the classical approach in scaling (see Anderson, 1981) and has been advocated by several authors in the field of structural equation modeling, most recently by Rosseel and Loh (2022), who also discuss its advantages.

In Experiment 1, we tested a quantitative belief-desire model of the intensity of disappointment and relief (Reisenzein, 2009) elicited by unobtained gains and losses in monetary lotteries. Belief and desire strengths were experimentally manipulated by varying, respectively, the objective probability and size of a possible monetary gain or loss (*cf.* Mellers et al., 1997). Nonlinear regression was used to fit the quantitative emotion models to the data of the individual participants, and the squared correlation between predicted and measured emotion intensity was used as the index of global model fit. For details, readers are referred to the original article (Junge and Reisenzein, 2013).

High fits of the emotion models were obtained for the indirect scales of most participants: $R^2$ was >0.90 for 68% of the participants if the MLDS scale values were used as the dependent variable, and for 90% if the AFM scale values were used. The explained variance in emotion intensity is so high that one may conclude that beliefs plus desires are sufficient causes of the intensity of relief and disappointment, as the tested emotion models assume. Furthermore, the pattern of scale values corresponded to the predicted pattern of a (nonlinear) fan for nearly all participants. In contrast, if emotion intensity ratings (the mean of two repeat measurements) were used as the dependent variable, only 13% of the participants attained an $R^2 > 0.90$ for relief and only 38% for disappointment. In addition, a separate test of the predicted linear interaction effect of the experimental manipulations on emotion intensity, reliably detected this interaction for the difference scales, but missed it for disappointment if the direct ratings were used. Incidentally, the better performance of the AFM scalings in this as well as the second study by Junge and Reisenzein (2013) might mean that GPCs contain more than just ordinal information about intensity differences.

In the second study, Junge and Reisenzein (2013, Experiment 2) tested a theory of (some) determinants of the intensity of disgust. Disgusting pictures were experimentally varied in size (big or small) and coloration (normal colored or false colored). Based on evolutionary considerations, it was predicted that the two manipulations would have an additive or superadditive effect on emotion intensity. Again, the difference scalings revealed the predicted pattern for the majority of the participants. For example, pooled across four experimentally manipulated disgust pictures, 51% of the participants conformed to the disgust model for the MLDS scalings and 85% for the AFM scalings, but only 30% did so for the ratings (made only once in this study, but after the GPC task).

These findings are important because they demonstrate the scientific utility of the indirect scaling methods. Experiment 1 showed that difference scalings of emotion intensity, but not direct intensity ratings, allowed to obtain support for quantitative emotion theories on the level of the individual subjects (Junge & Reisenzein, Study 1). Because most theories in psychology are formulated on the level of the individual, this is the level on which they should be preferably tested—a methodological recommendation repeatedly given (see, e.g., Estes, 1956; Woike et al., 2023) but still too rarely followed, particularly in emotion research. Experiment 2 demonstrated the same point for tests of ordinal causal hypotheses (Junge and Reisenzein, 2013, Exp. 2). Furthermore, the experiments demonstrated that difference scalings increase the power of statistical tests on both the individual and group levels. For example, they allowed to reliably detect predicted interaction effects, which are often missed with direct ratings (e.g., Nagengast et al., 2011).

## 4.4 Two reasons for the superior performance of difference scalings

One reason for the superior performance of the indirect scales compared to direct ratings in the reported tests of emotion models is their greater precision. This is in part simply a consequence of the fact that the indirect scales were based on a much larger set of judgments (although it should be noted that each GPC judgment provides only information about the difference between two emotion intensities). It could therefore be argued that, instead of using GPCs, one could simply replicate stimulus ratings more often and average them. This is standard practice in direct scalings of sensations of individual subjects, where the stimuli are presented numerous times (e.g., 50 times in Montgomery, 1982). However, apart from the fact that this does not address the limited resolution of ratings nor improve their scale level, numerous repeated ratings are usually not possible for affective stimuli (see also, Anderson, 1981). The main reason is that most emotional stimuli (e.g., affective pictures) are easy to memorize and participants could therefore simply reproduce their previous ratings. Aggregating ratings across participants to increase reliability is also of limited usefulness, because there are often large interindividual differences in emotional reactions to the same stimuli. Finally, the use of multiple indicators to increase the reliability of emotion ratings (e.g., Kline, 2016) is restricted, among other factors, by the fact that for many emotions, it is difficult to find more than a few emotion terms that have sufficient semantic similarity (e.g., what would be good multiple indicators for relief or disappointment?).

A second reason for the superior performance of the difference scales in our tests of emotion theories (Junge and Reisenzein, 2013) could have been that they approximated the metric scale level better than the ratings. This issue is addressed next.

# 5 Testing measurement axioms

As mentioned in the introduction, our approach to measurement combines the modern psychometric (i.e., latent-variable) approach to measurement, in our case represented by probabilistic difference scaling models, with the representational theory of measurement (RTM; e.g., Suppes and Zinnes, 1963; Krantz et al., 1971). This combination is facilitated by the fact that an axiomatic measurement theory for difference data—the data that constitute the input to the difference scaling models—exists (Krantz et al., 1971, Ch. 6). However, in our view, the integration of the latent-variable and RTM approaches to measurement requires a non-standard interpretation of RTM. To make clear where we differ from the standard interpretation of RTM, we briefly summarize it first.

## 5.1 The standard representation of RTM, illustrated for difference structures

The main goal of RTM is to specify the conditions, formulated as axioms, that the qualitative (typically, ordinal) relations among the levels of a variable must fulfill to allow a homomorphic (structure-preserving) mapping into a subset of the numbers, usually the reals. In the case of difference measurement, the qualitative (ordinal) structure is $<A \times A, \succsim>$ and the numerical structure is $<\mathbb{R}, \geq>$. For example, in difference measurement of emotion intensity, $A$ is a set of affective stimuli, $A \times A$ is the set of stimulus pairs $(a, b)$ from $A$, and $\succsim$ is the ordering of perceived differences in intensities of the feeling evoked by pairs of stimuli $(a, b)$ in a difference judgment task. The most direct way of obtaining these difference comparisons is the QC task (Section 3.1); however, they can also be retrieved from GPCs, as follows (Roberts, 1979; Orth, 1982): For all pairs of stimulus pairs $(a, b; c, d)$, $ab \succ cd$ (the intensity difference between the feelings elicited by $a$ and $b$ is greater than that between the feelings elicited $c$ and $d$) if GPC$(a, b) \succ$ GPC$(c, d)$ (example: $a$ is judged as eliciting *much more* pleasure than $b$, while $c$ is judged as eliciting *somewhat* more pleasure than $d$). If the two GPC judgments are equal, one is randomly chosen to be greater.

The axioms of difference structures impose constraints on the relation $\succsim$ which, when met, entail the existence of an interval-scale representation of the difference structure. That is, they entail the existence of a real-valued function $\Psi$ defined on $A$ that is unique up to a positive linear transformation, such that the biconditional (3) holds: (Krantz et al., 1971):

$$ab \succsim cd \text{ if, and only if } \Psi(a) - \Psi(b) \geq \Psi(c) - \Psi(d) \quad (3)$$

The two main testable axioms of difference structures in the standard axiomatization (Krantz et al., 1971) are the weak ordering axiom, and the axiom of weak monotonicity or the sextuple condition. The *weak ordering axiom* requires that $\succsim$ is a weak order (i.e., transitive and connected). It thus expresses the assumption, already discussed in Section 3.5, that people are able to consistently order intensity differences. The *sextuple axiom* is generally regarded as the central testable axiom of difference structures in the standard axiomatization (Krantz et al., 1971; Köbberling, 2006; see already Hölder, 1901). It is so called because it applies to sextuples of ordered stimuli $a \precsim b \precsim c$ and $a' \precsim b' \precsim c'$, for which it requires the condition (4) to hold:

$$\text{If } ab \succsim a'b' \text{ and } bc \succsim b'c' \text{ then } ac \succsim a'c' \quad (4)$$

For the ~ part of $\succsim$, axiom [4] reads: If $ab \sim a'b'$ and $bc \sim b'c'$, then $ac \sim a'c'$: If two adjoining intervals (judged intensity differences) $ab$ and $bc$ are equivalent in size to two other adjoining intervals $a'b'$ and $b'c'$, then the combined interval $ac$ is equivalent to $a'c'$ (for a graphical illustration see Krantz et al., 1971, p. 145). The complete sextuple axiom merely extends this requirement by replacing ~ with $\succsim$ (Krantz et al., 1971, p. 146). The sextuple axiom is an ordinal implication of the fact that intervals between numbers are additive: If two adjoining intervals on the number line, $x - y$ and $y - z$ are, respectively, identical to or greater than two other intervals $x' - y'$ and $y' - z'$, then the addition of the two intervals, $x - y + y - z = x - z$, is identical to (greater than) $x' - z'$. Additivity is the central condition that intensity intervals must meet, in addition to being weakly ordered, to allow an interval scale representation (Michell, 2012).

In alternative axiomatizations of difference structures, the sextuple axiom is replaced by a stronger requirement, the quadruple axiom (e.g., Debreu, 1958; Luce and Suppes, 1965; see also Köbberling, 2006), which requires: if $ab \succsim cd$, then $ac \succsim bd$. In our studies (Junge and Reisenzein, 2016), we tested this stronger axiom, partly to make up for the nontestability of the weak ordering axiom with GPCs (see Section 5.3). However, if the quadruple axiom is fulfilled, so is the sextuple axiom.

## 5.2 A realist and deductivist interpretation of RTM

The standard descriptions of RTM have been taken to imply by some authors (e.g., Borsboom, 2005) that RTM theorists interpret quantities *non-realistically* or *instrumentalistically*. That is, they regard the numerical representation of a qualitative structure (the scale $\Psi$) as an intervening variable that is useful as a compact summary of the ordinal relations in the data and as a device for making inferences, but does not refer to an independently existing quantity.

Furthermore, the standard descriptions of RTM suggest *a particular order of inquiry for the actual measurement process*. According to this order of inquiry, which can be called "inductivist" (and which is actually in tension with the otherwise deductive approach to measurement advocated by RTM theorists), the measurement process begins with the collection of a set of data for a qualitative relation structure, such as $<A \times A, \succsim>$ in the case of difference measurement. These data are next examined to determine whether they fulfill the axioms of the measurement structure. The actual measurement process, the estimation of scale values, is only performed in the third step (e.g., by applying a suitable nonmetric scaling method), and only if the second step has a positive outcome. This order of inquiry is nearly always followed in empirical applications of RTM (e.g., Schneider, 1982).

Although these interpretations of RTM undoubtedly reflect the views of some proponents of RTM, they are not shared by all (e.g., Orth, 1982; Westermann, 1983; Diéz Calzada, 2000). More importantly, the mathematical core of RTM—the qualitative relation structure, the representing numerical structure, the axioms, and the representation and uniqueness theorems derived from them—is

equally compatible with a realist interpretation of quantities, and a deductivist approach to axiom testing.

### 5.2.1 A realist interpretation of RTM

According to the realist view of quantities—that we endorse for at least some mental quantities including emotions—quantitative variables exist (or are hypothesized to exist) prior to and independent of any attempts to measure them, and the process of measurement is the attempt to determine the levels of the variable in a specific case (here and in part of what follows, we rely on Michell, 1999, 2005). As argued by Borsboom (2005), a realist view of quantities fits naturally with latent variable theories, to which ODS and MLDS belong.

As pointed out by Michell (1999), the concept of quantity (quantitative magnitude) was first defined in fully explicit and precise form by Hölder (1901, see Michell and Ernst, 1996, 1997) in his axioms of quantity. According to Hölder (1901), quantities are continuous variables whose levels are different degrees or gradings of a homogenous property, that stand to each other in a specific set of relations that together constitute an additive structure (Michell, 1999, 2005). Like the quantitative variable levels themselves, the relations between them may or may not be directly observable. In the latter case, which is characteristic for psychological quantities, what is observable—at least by the scientist—are only the manifestations or causal effects of the latent quantity in empirical measurements.

This realist view of latent quantities implies, among others, that the metric structure of the same latent variable (1) can manifest itself in somewhat different observable ways in the data resulting from different measurement procedures; (2) can get partly or completely lost in an attempted measurement process (e.g., O'Brien, 1985); and (3) that, as assumed in latent-variable measurement theories, measurements are always contaminated with some degree of error.

Furthermore, from a realist perspective, the assumptions (a) that a latent variable posited in a substantive theory (e.g., an emotion theory) is quantitative, and (b) that a particular measurement of this variable has a certain metric scale level (interval, ratio), are just two additional empirical assumptions made when testing the theory. The first assumption is implicitly made whenever a substantive theory postulates quantitative functional relations between variables, for these are only meaningful for quantitative variables. The second assumption is implicitly or explicitly made whenever researchers attempt to test the quantitative relations postulated in the theory by measuring their variables, for such tests are only meaningful if the measurements preserve (enough of) the variables' metric structure.

Although the "metricity" assumptions [a] and [b] are structural rather than causal (see Michell, 1999), they can, in principle, be tested like other theoretical assumptions; that is, by deriving testable consequences from them and then testing these consequences. Generally speaking, metricity assumptions have two kinds of testable implications. First, the substantive theory $\mathbf{T_s}$, together with an associated measurement theory $\mathbf{T_m}$ (these are linked by their reference to the same quantities), entail that the quantitative relations among the latent variables postulated in $\mathbf{T_s}$, will also be observed for the measurements of these variables up to the scale level of the measurements, and up to measurement error. Therefore, one can test the metricity assumptions, if indirectly and holistically, by testing the empirical predictions of the theory with a set of measurements that one simultaneously hypothesizes to be metric. This is the classical approach taken in tests of latent-variable structural equation models (e.g., Kline, 2016), where the causal model and the measurement model are simultaneously estimated. Essentially the same holistic test of measurement assumptions is advocated in Anderson's (Anderson, 1981, 1982) functional measurement method.

Second, $\mathbf{T_m}$ entails that the measurements of the latent variable fulfill, up to random error, the axioms of appropriate RTM measurement structures (see 5.2.2). This test of metricity is independent of $\mathbf{T_s}$ and therefore more diagnostic. However, analogous to the holistic test of metricity assumptions, a realist interpretation of latent quantities suggests a deductive rather than inductive order of inquiry when testing measurement axioms.

### 5.2.2 A deductivist order of inquiry for testing measurement axioms

The deductivist order of inquiry in the measurement process has been elaborated in a series of papers by Westermann (1982, 1983, 1985). It begins with a proposed numerical measurement of a latent variable (e.g., scale values estimated by ODS) and only subsequently tests whether the scale values fulfill the axioms of an appropriate measurement structure (a closely related approach was proposed by Orth, 1982). In the context of the probabilistic difference scaling models, the deductive test of measurement axioms appears as just another diagnostic test, performed after the scaling, of the assumptions underlying the scaling model (see Maloney and Yang, 2003; Knoblauch and Maloney, 2008). A major benefit of testing measurement axioms in the context of probabilistic difference scaling models is that doing so provides a solution to a long-standing problem of RTM (see Krantz et al., 1971; Luce et al., 1990), the problem of accounting for measurement errors: Because ODS and MLDS are probabilistic latent variable models, they automatically yield an estimate of judgment error that can be used to construct a statistical test of axiom adherence (see Section 5.3).

Note, however, that the deductive order of inquiry for testing measurement axioms suggests an important modification regarding *how*, precisely, measurement axioms are tested (Junge and Reisenzein, 2016). Generally speaking, a measurement axiom is tested by selecting cases that fulfill the antecedent (if) condition of the axiom, and then checking whether these cases also fulfill the consequens (then) part of the axiom. In the classical RTM approach, this test, illustrated for the sextuple axiom, is implemented as follows: One selects sextuples of stimuli $(a, b, c, a', b', c')$ from $A$ in $<A \times A, \gtrsim>$ that fulfill the condition $ab \gtrsim a'b'$ and $bc \gtrsim b'c'$, and then checks whether these sextuples also fulfill $ac \gtrsim a'c'$.

However, if the order of inquiry begins with actual (proposed) numerical measurements, it is only consequential, as well consistent with the general deductive approach to theory testing, to use the *estimated scale values* to select the antecedent cases of the axiom. The reason is that the scale values are the best available estimates of the latent variable values, and much less contaminated by error than is each individual comparative judgment (which is usually only made once). Hence, the deductivist approach suggests the following modification of the axiom test in ODS and MLDS (Junge and Reisenzein, 2016): The test cases are not chosen by relying on $\gtrsim$ (for the sextuple axiom, by selecting sextuples of stimuli that fulfill the condition $ab \gtrsim a'b'$ and $bc \gtrsim b'c'$), but by selecting sextuples for which $\Psi(a) - \Psi(b) \geq \Psi(a') - \Psi(b')$ and $\Psi(b) - \Psi(c) \geq \Psi(b') - \Psi(c')$. For these sextuples, one then checks whether $ac \gtrsim a'c'$ is fulfilled in the empirical difference data (Junge and Reisenzein, 2016).

## 5.3 Testing the quadruple axiom

### 5.3.1 The test procedure

As explained in Section 5.1, the two main testable axioms of difference structures are the weak ordering axiom and the sextuple axiom (or, in a different axiomatization, the stronger quadruple axiom). In our study on axiom adherence (Junge and Reisenzein, 2016) we could not test the weak ordering axiom, because this axiom is necessarily fulfilled if difference comparisons are derived from GPCs (see Orth, 1982; Junge and Reisenzein, 2016). However, as argued in Section 3.5, the assumption hat people can order differences of emotion intensity is intuitively plausible and there is evidence from difference scaling studies of sensations and perceptions that this axiom is usually fulfilled (up to random error). The focus of Junge and Reisenzein (2016) was therefore on the test of the quadruple axiom, which, as mentioned, implies the sextuple condition.

To test the quadruple axiom, we used a modified version of a parametric bootstrap test proposed by Maloney and Yang (2003) and Knoblauch and Maloney (2008) for testing axiom violation in the context of MLDS. This test was adapted to account for the fact that we used GPCs rather than QCs, meaning that the scale values and error variance were estimated by ODS rather than MLDS, and that the difference comparisons (*ab*; *cd*) were derived from the GPCs. Also different from Maloney and Yang (2003), we used a traditional performance criterion, the percentage of axiom adherence (= 100 – percent of axiom violations) as the test statistic. Most important, for reasons explained above, we used the estimated scale values instead of the participant's ordinal judgments to select the test cases for the quadruple test.

Concretely, the axiom test was as follows. In the first step, the scale values estimated by ODS were used to select quadruples (*a*, *b*; *c*, *d*) that fulfilled the antecedent condition of the quadruple axiom. To account for the fact that participants cannot discriminate differences if they are too small, a conservative discriminability threshold was set. Furthermore, we selected only quadruples for which $|\Psi_a - \Psi_b| > |\Psi_c - \Psi_d|$ (Orth, 1982) to account for the fact that small discriminable differences, that might still be detected in direct difference comparisons, cannot reveal themselves in GPCs because of the limited resolution of the response scale.

In the second step, the scale values and error variance of the judgments estimated by ODS were used to generate 10.000 simulated GPC responses, which were expanded to QCs. These simulated responses reflect the performance of an "ideal observer" (Maloney and Yang, 2003), i.e., a hypothetical twin of the participant who judges each quadruple according to the ODS model, given the participant's scale values and error variance. From these simulated QCs, the ideal observer's response to the antecedent of the quadruple axiom was extracted for the test cases of the axiom. Hence, the actual form of the tested axiom was: If $|\Psi_a - \Psi_b| > |\Psi_c - \Psi_d|$ then $ac \succ bd$.

In the third step, the percentage of correct responses to the test cases of the axiom (i.e., responses where $ac \succ bd$) was computed for each simulation, and this performance index was accumulated into a bootstrap distribution. This distribution reflects the variability of the responses of the ideal observer who responds repeatedly to the axiom test cases. Finally, the percentage of correct responses of the participant was compared to the bootstrap distribution. If the probability of the obtained percentage correct was < 0.05, we concluded that the participant systematically violated the quadruple axiom. Otherwise, we concluded that the null hypothesis—the participant responded in accordance with the quadruple axiom—can be retained.

### 5.3.2 Results

For the six emotions investigated by Junge and Reisenzein (2016), the hypothesis that the participants' ODS scale values adhered to the quadruple axiom could be retained for most participants: amusement 71%; relief 74%; disgust 81%; surprise 88%; pleasantness 97%, and disappointment 97%. These findings suggest that the ODS scale values of most participants were metric or more precisely, interval-scaled. If one grants that the natural zero point of emotion intensity (the absence of emotion) was, with acceptable precision, estimated by the simultaneously collected direct intensity ratings, a ratio scale can be obtained for the axiom-conforming participants by linearly transforming their ODS values into the range of their intensity ratings (see Section 2.9).

## 5.4 Testing the metricity of direct scalings of emotion intensity

### 5.4.1 The test procedure

If one accepts that the ODS scale of participants who passed the quadruple test is metric, one has a standard of comparison for deciding whether the direct emotion intensity scalings of these participants are metric as well. The underlying logic is this: If the emotion intensities estimated by ODS are interval-scaled, then any other interval-scale measurement *M* of the same emotion intensities is a linear transformation of the ODS scale and should therefore be linearly correlated with the ODS scale as highly as the reliability of the ODS scale and *M* permit. Based on this logic, Junge and Reisenzein (2016) constructed another bootstrap test to test the metricity of the direct emotion ratings. In this test, the ODS scalings were treated as error-free (which they nearly were), whereas the error contained in the ratings was estimated from the ratings' re-test reliability (see Junge and Reisenzein, 2016).

For each participant and emotion, 10.000 simulated ratings were generated from the ODS scale by perturbing the scale values with normal error corresponding to that of the ratings. This procedure simulates a hypothetical twin of the participant who uses the ODS scale values to make the ratings, but makes random errors corresponding to the error level of the ratings. Each simulated set of ratings was then linearly correlated with the ODS scale values, and the correlations were accumulated into a bootstrap distribution. This distribution reflects the expected variability of the correlation between the direct and the ODS scale for a person who operates with the ODS scale values, but makes random errors in the ratings corresponding to the ratings' error level. Finally, the bootstrap distribution was compared to the actual correlation between the direct and indirect scales obtained for the participant.

### 5.4.2 Results

In Study 1 of Junge and Reisenzein (2016), 44% of the participants whose ODS scale values for pleasure were metric according to the quadruple test, and 23% of those whose ODS scale values of disgust

were metric, also passed the metricity test for the corresponding ratings. Similar findings were obtained in Study 2 for ratings of amusement and surprise evoked by quiz items, and in Study 3 for ratings of disappointment about unobtained gains, and of relief about unobtained losses, in monetary lotteries. Hence, for all six investigated emotions, the direct ratings of emotion intensity of the majority of the participants deviated statistically significantly from the ODS scale values.

Notwithstanding the significant deviations from the metric (interval) scale level, it is reasonable to ask: Did the obtained direct ratings of emotion intensity at least *approximate* the linear ODS scale? A rough answer to this question is suggested by the size of the linear correlation between the direct and indirect scales of the participants who passed the quadruple test. In Study 1, this correlation was on average 0.80 for pleasure and 0.81 for disgust, although with a wide range (0.43 to 0.92 for pleasure and 0.18 to 0.96 for disgust). Similar correlations were obtained in Study 2 for surprise ($M = 0.86$, *range* = 0.67 to 0.94) and amusement ($M = 0.88$, *range* = 0.52 to 0.98) and in Study 3 for relief ($M = 0.78$, *range* = −0.18 to 0.96) and disappointment ($M = 0.80$, *range* = −0.36 to 0.96). Judged by traditional psychometric standards, the average obtained correlation of 0.82 would be considered fair. Thus, despite the statistically significant deviations of the emotion ratings of most participants from the interval scale level, the majority seemed to approximate linearity to a fair degree. This conclusion supports the assumption (e.g., Anderson, 1981, 1982) that the response function of carefully constructed rating scales is approximately linear. Although far from perfect ($R^2 = 0.67$), the found degree of approximation of the ratings to the linear scale (represented by the ODS scale) may be sufficient for some kinds of analyses. However, as demonstrated by the results of Junge and Reisenzein (2013), emotion ratings are not precise enough and/or not close enough to metric to support tests of emotion theories on the individual subject level.

## 6 When can and should difference scaling be used?

Although we have focused on emotional experiences in this article, the proposed measurement approach can also be used to measure the intensity of sensations, bodily feelings, and other mental states characterized by an experiential quality of varying intensity. As mentioned, applications of difference scaling methods in both the older (e.g., Orth, 1982; Schneider, 1982) and more recent psychological literature (e.g., Boschman, 2001; Maloney and Yang, 2003; Maloney and Knoblauch, 2020) found that these methods yield precise measurements on an interval scale level for a variety of sensations and perceptions. Regarding the measurement of yet other mental states, particularly those whose conceptualization as quantities is a priori uncertain, caution is indicated (see Michell, 2012); in these cases, the proposed deductive method of testing measurement axioms could help to clarify the situation.

Despite the advantages of difference scaling methods, specifically ODS, for measuring the intensity of emotions, they are not the method of choice in all situations. This is so for two main reasons (see also, Junge and Reisenzein, 2013). First, like other indirect scaling methods, difference scaling cannot be used in all measurement contexts. In particular, it cannot be used when it is not possible or meaningful to compare multiple affective stimuli, or to present them repeatedly in GPCs or QCs. This is often the case in real-life situations (e.g., emotional reactions to outcome of exams; Pekrun and Bühner, 2014). Even in the laboratory, repeated stimulus comparisons are problematic for stimuli such as tastes and smells (Cardello, 2017).

Second, even when difference scaling methods are applicable, they are—again like other indirect measurement methods—more costly than direct scaling methods in terms of the time, effort and resources required for data collection and the calculation of scale values (Cardello and Jaeger, 2010; Cardello, 2017). However, it should be noted that these costs can be substantially reduced through computerized stimulus presentation, data collection, and scale value estimation (see Knoblauch and Maloney, 2008; Junge and Reisenzein, 2013). Although a time disadvantage in data collection remains, it is in fact not very large for ODS with up to about 12 stimuli, especially if the alternative consists of direct scalings repeated once (to increase reliability). For example, with 10 stimuli, there are 45 possible GPCs, but it appears that this number can be reduced by half without significantly degrading the scale value estimates (Boschman, 2001). This results in a comparable number of judgments to those needed for once-repeated, direct stimulus ratings. For 12 stimuli, the choice is between 24 ratings and about 30 GPCs. Furthermore, the time required to complete a GPC judgment is similar to that needed for a rating, and GPCs seem to be no more difficult to make than ratings. However, one potentially important difference remains: GPCs require twice as many stimulus presentations (2 in each trial) than direct scalings.

Whether the additional costs of difference scaling methods—even those of the economical ODS method—are an acceptable trade-off for obtaining more precise, less biased, and closely metric measurements, depends, among other factors, on the research question. Difference scaling methods are likely most useful in basic research when high-precision, metric measurements are desired to test substantive theories, particularly quantitative theories tested at the level of the individual. In contrast, in applied settings, where time constraints are often a preeminent concern, or when less precise and only roughly metric measurements are sufficient, difference scaling methods can be inefficient, i.e., too costly for the additional information they provide. In these situations, as well as in settings where difference scaling cannot be used (see above), optimized versions of the classical rating scale (see Anderson, 1982), or the newer labeled affective magnitude scales mentioned in Section 2.2.3, are currently (still) the best alternatives. And in some research contexts, ordinal or even qualitative (presence/absence) assessments of emotion will do.

Finally, even if the intensity of emotions is measured by ratings or other direct scaling methods, difference scalings are useful for checking the scale level obtained with these methods (Westermann, 1983).

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: no new data are reported. The datasets of our

original, published studies referred to in the article will be made available to qualified researchers. Requests to access these datasets should be directed to rainer.reisenzein@uni-greifswald.de.

## Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

## Author contributions

RR: Writing – original draft, Writing – review & editing. MJ: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *J. Royal Stat. Soc.* 41, 287–297.

Agresti, A. (2010). *Analysis of ordinal categorical data. 2nd* Edn. Hoboken: Wiley.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., and Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, N. H. (1981). *Foundations of information integration theory*. Cambridge, MA: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory*. Cambridge, MA: Academic Press.

Ares, G., and Vidal, L. (2020). "Measuring liking for food and drink" in *Handbook of eating and drinking: Interdisciplinary perspectives*. ed. H. L. Meiselman (London: Springer Nature), 235–256.

Bartoshuk, L. (2014). The measurement of pleasure and pain. *Perspect. Psychol. Sci.* 9, 91–93. doi: 10.1177/1745691613512660

Bartoshuk, L. M., Fast, K., and Snyder, D. J. (2005). Differences in our sensory worlds: invalid comparisons with labeled scales. *Curr. Dir. Psychol. Sci.* 14, 122–125. doi: 10.1111/j.0963-7214.2005.00346.x

Baumgartner, H., and Steenkamp, J. B. E. (2001). Response styles in marketing research: a cross-national investigation. *J. Mark. Res.* 38, 143–156. doi: 10.1509/jmkr.38.2.143.18840

Beck, J. (2015). Analogue magnitude representations: a philosophical introduction. *Br. J. Philos. Sci.* 66, 829–855. doi: 10.1093/bjps/axu014

Birnbaum, M. H. (1982). "Problems with so-called "direct" scaling" in *Selected sensory methods: Problems and approaches to measuring hedonics*. eds. J. T. Kuznicki, A. F. Rutkiewic and R. A. Hedges (West Conshohocken, PA: ASTM International).

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: identifying the scale origin. *Psychol. Methods* 9, 453–465. doi: 10.1037/1082-989X.9.4.453

Böckenholt, U. (2006). Thurstonian-based analyses: past, present, and future utilities. *Psychometrika* 71, 615–629. doi: 10.1007/s11336-006-1598-5

Borg, G. (1962). *Physical performance and perceived exertion*. Lund: Gleerup.

Borsboom, D. (2005). *Measuring the mind. Conceptual issues in contemporary psychometrics*. New York: Cambridge University Press.

Boschman, M. C. (2001). DifScal: a tool for analyzing difference ratings on an ordinal category scale. *Behav. Res. Methods Instruments Comput.* 33, 10–20. doi: 10.3758/BF03195343

Brown, A., and Maydeu-Olivares, A. (2018). "Modeling forced-choice response formats" in *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. eds. P. Irwing, T. Booth and D. Hughes (London: John Wiley & Sons), 523–569.

Bürkner, P. C. (2017). brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Cardello, A. V. (2017). Hedonic scaling: assumptions, contexts and frames of reference. *Curr. Opin. Food Sci.* 15, 14–21. doi: 10.1016/j.cofs.2017.05.002

Cardello, A. V., and Jaeger, S. R. (2010). "Hedonic measurement for product development: new methods for direct and indirect scaling" in *Consumer-driven innovation in food and personal care products*. eds. S. R. Jaeger and H. MacFie (Cambridge: Woodhead Publishing), 135–174.

Carnap, R. (1966). *Philosophical foundations of physics*. New York: Basic Books.

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package ordinal. Available at: https://cran.r-hub.io/web/packages/ordinal/vignettes/clm_article.pdf

Cohn, J. (1894). Experimentelle Untersuchungen über die Gefühlsbetonung der Farben, Helligkeiten und ihrer Combination. *Philos. Stud.* 10, 562–604.

Debreu, G. (1958). Stochastic choice and cardinal utility. *Econometrica* 26, 440–444. doi: 10.2307/1907622

Diéz Calzada, J. A. (2000). "Structuralist analysis of theories of fundamental measurement" in *Structuralist knowledge representation: Paradigmatic examples*. eds. W. Balzer, J. Sneed and C. U. Moulines (Amsterdam: Rodopi), 19–49.

Ekman, P. (1992). An argument for basic emotions. *Cognit. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Ellermeier, W., and Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Percept. Psychophys.* 62, 1505–1511. doi: 10.3758/BF03212151

Engen, T. (1971). "Psychophysics II. Scaling methods" in *Woodworth & Schlosberg's experimental psychology*. eds. J. W. Kling and L. A. Riggs. *3rd* ed (New York: Holt, Rinehart & Winston).

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychol. Bull.* 53, 134–140. doi: 10.1037/h0045156

Farkas, K., Green, E., Rigby, D., Cross, P., Tyrrel, S., Malham, S. K., et al. (2021). Investigating awareness, fear and control associated with norovirus and other pathogens and pollutants using best–worst scaling. *Sci. Rep.* 11:11194.

Farquhar, P. H., and Keller, L. R. (1989). Preference intensity measurement. *Ann. Oper. Res.* 19, 205–217. doi: 10.1007/BF02283521

Fechner, G. T. (1860). *Elemente der Psychophysik [Elements of psychophysics]*. Leipzig: Breitkopf u. Härtel.

Fechner, T. (1871). *Zur experimentalen Ästhetik [On experimental aesthetics]*. Leipzig: Hirzel.

Finn, A., and Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: the case of food safety. *J. Public Policy Mark.* 11, 12–25. doi: 10.1177/074391569201100202

Galanter, E. (1990). Utility scales of monetary and nonmonetary events. *Am. J. Psychol.* 103, 449–470. doi: 10.2307/1423318

García-Pérez, M. A., and Alcalá-Quintana, R. (2012). On the discrepant results in synchrony judgment and temporal-order judgment tasks: a quantitative model. *Psychon. Bull. Rev.* 19, 820–846. doi: 10.3758/s13423-012-0278-y

Gelman, A., and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Greene, W. H., and Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge: Cambridge University Press.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Haghiri, S., Wichmann, F. A., and von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *J. Vis.* 20:14. doi: 10.1167/jov.20.9.14

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken, NJ: John Wiley & Sons Inc.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. *Berichte über die Verhandlungen der Königlichen Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* 53, 1–64.

Jaeger, S. R., and Cardello, A. V. (2009). Direct and indirect hedonic scaling methods: a comparison of the labeled affective magnitude (LAM) scale and best–worst scaling. *Food Qual. Prefer.* 20, 249–258. doi: 10.1016/j.foodqual.2008.10.005

Jaeger, S. R., Jørgensen, A. S., Aaslyng, M. D., and Bredie, W. L. P. (2008). Best-worst scaling: an introduction and initial comparison with monadic rating for preference elicitation with food products. *Food Qual. Prefer.* 19, 579–588. doi: 10.1016/j.foodqual.2008.03.002

Junge, M., and Reisenzein, R. (2013). Indirect scaling methods for testing quantitative emotion theories. *Cognit. Emot.* 27, 1247–1275. doi: 10.1080/02699931.2013.782267

Junge, M., and Reisenzein, R. (2015). Maximum likelihood difference scaling versus ordinal difference scaling of emotion intensity: a comparison. *Qual. Quant.* 49, 2169–2185. doi: 10.1007/s11135-014-0100-1

Junge, M., and Reisenzein, R. (2016). Metric scales for emotion measurement. *Psychol. Test Assess. Model.* 58, 497–530.

Kim, K. O., and O'Mahony, M. (1998). A new approach to category scales of intensity I: traditional versus rank-rating. *J. Sens. Stud.* 13, 241–249. doi: 10.1111/j.1745-459X.1998.tb00086.x

Kingdom, F. A. A., and Prins, N. (2010). *Psychophysics: A practical introduction*. London: Elsevier.

Kline, R. B. (2016). *Principles and practice of structural equation modeling. 4th* Edn. New York: Guilford Press.

Knoblauch, K., and Maloney, L. T. (2008). MLDS: maximum likelihood difference scaling in R. *J. Stat. Softw.* 25, 1–28. doi: 10.18637/jss.v025.i02

Köbberling, V. (2006). Strength of preference and cardinal utility. *Economic Theory* 27, 375–391. doi: 10.1007/s00199-005-0598-5

Kosmidis, I. (2014). Improved estimation in cumulative link models. *J. R. Stat. Soc. Ser. B* 76, 169–196. doi: 10.1111/rssb.12025

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971). *Foundations of measurement. I: Additive and polynomial representations*. New York: Academic Press.

Kron, A. (2019). Rethinking the principles of emotion taxonomy. *Emot. Rev.* 11, 226–233. doi: 10.1177/1754073919843185

Külpe, O. (1893). *Grundriss der Psychologie auf experimenteller Grundlage [Outlines of psychology]*. Leipzig: Engelmann.

Laming, J., and Laming, D. (1996). J. Plateau: On the measurement of physical sensations and on the law which links the intensity of these sensations to the intensity of the source; J. Plateau: Report on 'psychophysical study: theoretical and experimental research on the measurement of sensations, particularly sensations of light and of fatigue' by Mr. Delboeuf. *Psychol. Res.* 59, 134–144.

Laucken, U. (1974). *Naive Verhaltenstheorie [The folk theory of behavior]*. Stuttgart: Klett.

Liddell, T. M., and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: what could possibly go wrong? *J. Exp. Soc. Psychol.* 79, 328–348. doi: 10.1016/j.jesp.2018.08.009

Lim, J. (2011). Hedonic scaling: a review of methods and theory. *Food Qual. Prefer.* 22, 733–747. doi: 10.1016/j.foodqual.2011.05.008

Lishner, D. A., Cooter, A. B., and Zald, D. H. (2008). Addressing measurement limitations in affective rating scales: development of an empirical valence scale. *Cognit. Emot.* 22, 180–192. doi: 10.1080/02699930701319139

Lorr, M. (1989). "Models and methods for measurement of mood" in *The measurement of emotions*. ed. R. Plutchik (Cambridge, MA: Academic Press), 37–53.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge: Cambridge University Press.

Luce, R. D. (2010). Interpersonal comparisons of utility for 2 of 3 types of people. *Theor. Decis.* 68, 5–24. doi: 10.1007/s11238-009-9138-2

Luce, R. D., Krantz, D. H., Suppes, P., and Tversky, A. (1990). *Foundations of measurement Vol 3: Representation, axiomatization, and invariance*. London: Academic Press.

Luce, R. D., and Suppes, P. (1965). "Preference, utility, and subjective probability" in *Handbook of mathematical psychology*. eds. R. D. Luce, R. R. Bush and E. Galanter, vol. *III* (New York: Wiley), 252–410.

Machina, M. J. (1981). "Rational" decision making versus "rational" decision modelling? Review of Maurice Allais and ole Hagen (Eds.). Expected utility hypotheses and the Allais paradox: contemporary discussions of decisions under uncertainty with Allais' rejoinder (Theory and Decision library, Vol. 21). *J. Math. Psychol.* 24, 163–175. doi: 10.1016/0022-2496(81)90041-9

Major, D. R. (1895). On the affective tone of simple sense-impressions. *Am. J. Psychol.* 7, 57–77. doi: 10.2307/1412037

Maloney, L. T., and Knoblauch, K. (2020). Measuring and modeling visual appearance. *Ann. Rev. Vision Sci.* 6, 519–537. doi: 10.1146/annurev-vision-030320-041152

Maloney, L. T., and Yang, J. N. (2003). Maximum likelihood difference scaling. *J. Vis.* 3, 573–585. doi: 10.1167/3.8.5

Marks, L. E., Borg, G., and Ljunggren, G. (1983). Individual differences in perceived exertion assessed by two new methods. *Percept. Psychophys.* 34, 280–288. doi: 10.3758/BF03202957

Marks, L. E., and Gescheider, G. A. (2002). "Psychophysical scaling" in *Stevens' handbook of experimental psychology: Methodology in experimental psychology*. eds. H. Pashler and J. Wixted, vol. 4, 91–138.

Marley, A. A., and Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *J. Math. Psychol.* 49, 464–480. doi: 10.1016/j.jmp.2005.05.003

Masin, S. C. (2022). Old and new views on ratio judgment. In Fechner Day 2022: Proceedings of the 38th Annual Meeting of the International Society for Psychophysics, Lund, Sweden (pp. 61–66).

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cognit. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

McKelvey, R., and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* 4, 103–120. doi: 10.1080/0022250X.1975.9989847

Mellers, B. A., Schwartz, A., Ho, K., and Ritov, I. (1997). Decision affect theory: emotional reactions to the outcomes of risky options. *Psychol. Sci.* 8, 423–429. doi: 10.1111/j.1467-9280.1997.tb00455.x

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*, vol. *53*. Cambridge: Cambridge University Press.

Michell, J. (2005). The logic of measurement: a realist overview. *Measurement* 38, 285–294. doi: 10.1016/j.measurement.2005.09.004

Michell, J. (2012). "The constantly recurring argument": inferring quantity from order. *Theory Psychol.* 22, 255–271. doi: 10.1177/0959354311434656

Michell, J., and Ernst, C. (1996). The axioms of quantity and the theory of measurement. Translated from part I of Otto Hölder's german text "Die Axiome der Quantität und die Lehre vom Mass". *J. Math. Psychol.* 40, 235–252. doi: 10.1006/jmps.1996.0023

Michell, J., and Ernst, C. (1997). The axioms of quantity and the theory of measurement. Translated from part II of Otto Hölder's german text "Die Axiome der Quantität und die Lehre vom Mass". *J. Math. Psychol.* 41, 345–356. doi: 10.1006/jmps.1997.1178

Mielby, L. H., Edelenbos, M., and Thybo, A. K. (2012). Comparison of rating, best–worst scaling, and adolescents' real choices of snacks. *Food Qual. Prefer.* 25, 140–147. doi: 10.1016/j.foodqual.2012.02.007

Mohammad, S., and Bravo-Marquez, F. (2017). Emotion intensities in tweets. In Proceedings of the 6th joint conference on lexical and computational semantics (*SEM 2017) (pp. 65–77). Vancouver, Canada.

Montgomery, H. (1982). "Intra- and interindividual variations in the form of psychophysical scales" in *Social attitudes and psychophysical measurement*. ed. B. Wegener (Hillsdale, NJ: Erlbaum), 339–357.

Moscati, I. (2019). *Measuring utility: From the marginal revolution to behavioral economics*. Oxford: Oxford University Press.

Moskowitz, H. R., and Sidel, J. L. (1971). Magnitude and hedonic scales of food acceptability. *J. Food Sci.* 36, 677–680. doi: 10.1111/j.1365-2621.1971.tb15160.x

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K. T., and Trautwein, U. (2011). Who took the "×" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychol. Sci.* 22, 1058–1066. doi: 10.1177/0956797611415540

Nowlis, V., and Nowlis, H. H. (1956). The description and analysis of mood. *Ann. New York Acad. Sci.* 65:345. doi: 10.1111/j.1749-6632.1956.tb49644.x

O'Brien, R. M. (1985). The relationship between ordinal measures and their underlying values: why all the disagreement? *Qual. Quant.* 19, 265–277. doi: 10.1007/BF00170998

Orth, B. (1982). "A theoretical and empirical study of scale properties of magnitude-estimation and category rating scales" in *Social attitudes and psychophysical measurement*. ed. B. Wegener (Hillsdale, NJ: Erlbaum), 351–377.

Ortony, A. (2022). Are all "basic emotions" emotions? A problem for the (basic) emotions construct. *Perspect. Psychol. Sci.* 17, 41–61. doi: 10.1177/1745691620985415

Ortony, A., Clore, G. L., and Collins, A. (1988). *The cognitive structure of emotions*. Cambridge: University Press.

Pekrun, R., and Bühner, M. (2014). "Self-report measures of academic emotions" in *International handbook of emotions in education*. eds. R. Pekrun and L. Linnenbrink-Garcia (New York: Taylor & Francis), 561–579.

Peryam, D. R., and Pilgrim, F. J. (1957). Hedonic scale method of measuring food preferences. *Food Technol.* 11, 9–14.

Petrov, A. A., and Anderson, J. R. (2005). The dynamics of scaling: a memory-based anchor model of category rating and absolute identification. *Psychol. Rev.* 112, 383–416. doi: 10.1037/0033-295X.112.2.383

Plateau, J. (1872). Sur la mesure des sensations physique, et sur la loi qui lie l'intensité de ces sensations à l'intensité de la cause excitante. *Bulletins de l'académie royale des sciences, des lettres et des beaux-arts de Belgique* 33, 376–388.

Plutchik, R. (1989). "Measuring emotions and their derivatives" in *The measurement of emotions*. ed. R. Plutchik (Cambridge, MA: Academic Press), 1–35.

Poulton, E. C. (1989). *Bias in quantifying judgments*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

R Core Team (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

Reisenzein, R. (2009). Emotions as metarepresentational states of mind: naturalizing the belief–desire theory of emotion. *Cogn. Syst. Res.* 10, 6–20. doi: 10.1016/j.cogsys.2008.03.001

Reisenzein, R. (2012). "What is an emotion in the belief-desire theory of emotion?" in *The goals of cognition: Essays in honor of Cristiano Castelfranchi*. eds. F. Paglieri, L. Tummolini, R. Falcone and M. Miceli (Suwanee: College Publications), 181–211.

Reisenzein, R., and Franikowski, P. (2019). Improving theory tests by improving measurement: A test of the attributional theory of help-giving using ordinal difference scaling. Unpublished manuscript. Greifswald: University of Greifswald.

Reisenzein, R., and Franikowski, P. (2022). On the latency of object recognition and affect: evidence from temporal order and simultaneity judgments. *J. Exp. Psychol. Gen.* 151, 3060–3081. doi: 10.1037/xge0001244

Reisenzein, R., Hildebrandt, A., and Weber, H. (2020). "Personality and emotion" in *The Cambridge handbook of personality psychology*. eds. P. J. Corr and G. Matthews. *2nd* ed (Cambridge: Cambridge University Press), 81–99.

Reisenzein, R., Horstmann, G., and Schützwohl, A. (2019). The cognitive-evolutionary model of surprise: a review of the evidence. *Top. Cogn. Sci.* 11, 50–74. doi: 10.1111/tops.12292

Reisenzein, R., and Junge, J. (2024). *Ordinal difference scaling of hope followed by disappointment, and of fear followed by relief, in hypothetical scenarios*. University of Greifswald: Unpublished data.

Reisenzein, R., Junge, M., Studtmann, M., and Huber, O. (2014). "Observational approaches to the measurement of emotions" in *International handbook of emotions in education*. eds. R. Pekrun and L. Linnenbrink-Garcia (London: Taylor & Francis / Routledge), 580–606.

Reisenzein, R. (2015). "A short history of psychological perspectives on emotion" in *Oxford handbook of affective computing*. eds. R. A. Calvo, S. K. D'Mello, J. Gratch and A. Kappas (Oxford: Oxford University Press), 21–37.

Roberts, F. S. (1979). *Measurement theory: With applications to decision making, utility, and the social sciences*. Reading, Mass: Addison-Wesley.

Rosseel, Y., and Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological methods*. Advance online publication. doi: 10.1037/met0000503

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145–172. doi: 10.1037/0033-295X.110.1.145

Schifferstein, H. N. (2012). Labeled magnitude scales: a critical review. *Food Qual. Prefer.* 26, 151–158. doi: 10.1016/j.foodqual.2012.04.016

Schimmack, U., Böckenholt, U., and Reisenzein, R. (2002). Response styles in affect ratings: making a mountain out of a molehill. *J. Pers. Assess.* 78, 461–483. doi: 10.1207/S15327752JPA7803_06

Schneider, B. (1982). "The nonmetric analysis of difference judgments in social psychophysics: scale validity and dimensionality" in *Social attitudes and psychophysical measurement*. ed. B. Wegener (Hillsdale, NJ: Erlbaum), 317–337.

Schuster, A. L., Crossnohere, N. L., Campoamor, N. B., Hollin, I. L., and Bridges, J. F. (2024). The rise of best-worst scaling for prioritization: a transdisciplinary literature review. *J. Choice Model.* 50:100466. doi: 10.1016/j.jocm.2023.100466

Schutz, H. G., and Cardello, A. V. (2001). A labeled affective magnitude (LAM) scale for assessing food liking/disliking. *J. Sens. Stud.* 16, 117–159. doi: 10.1111/j.1745-459X.2001.tb00293.x

Sixtl, F. (1982). *Messmethoden der Psychologie: Theoretische Grundlagen und Probleme [Measurement methods of psychology: Theoretical foundations and problems]. 2nd* Edn. Weinheim/Basel: Beltz-Verlag.

Stevens, S. S. (1975). *Psychophysics*. New York: Wiley.

Stevens, S. S., and Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *J. Exp. Psychol.* 54, 377–411. doi: 10.1037/h0043680

Sullivan, R. (1971). Magnitude estimation and relative aversiveness of anxiety: phobia. *J. Abnorm. Psychol.* 78, 266–271. doi: 10.1037/h0031993

Suppes, P., and Zinnes, J. L. (1963). "Basic measurement theory" in *Handbook of mathematical psychology*. eds. R. D. Luce, R. R. Bush and E. Galanter, vol. *1* (New York: Wiley), 1–76.

Tal, E. (2020). Measurement in science. The Stanford encyclopedia of philosophy. Available at: https://plato.stanford.edu/archives/fall2020/entries/measurement-science/

Teghtsoonian, R., and Frost, R. O. (1982). The effects of viewing distance on fear of snakes. *J. Behav. Ther. Exp. Psychiatry* 13, 181–190. doi: 10.1016/0005-7916(82)90002-7

Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/h0070288

Titchener, E. B. (1902). Ein Versuch die Methode der paarweisen Vergleichung auf die verschiedenen Gefühlsrichtungen anzuwenden [An attempt to apply the method of paired comparisons to the different directions of feeling]. *Philos. Stud.* 20, 382–406.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response. *J. Math. Psychol.* 30, 306–316. doi: 10.1016/0022-2496(86)90034-9

Uher, J. (2018). Quantitative data from rating scales: an epistemological and methodological enquiry. *Front. Psychol.* 9:2599. doi: 10.3389/fpsyg.2018.02599

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Soc. Personal. Psychol. Compass* 17:e12740. doi: 10.1111/spc3.12740

von Winterfeldt, D., and Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Westermann, R. (1982). Empirical test of scale type resulting from the power law for heaviness. *Percept. Mot. Skills* 55, 1167–1173. doi: 10.2466/pms.1982.55.3f.1167

Westermann, R. (1983). Interval-scale measurement of attitudes: some theoretical conditions and empirical testing methods. *Br. J. Math. Stat. Psychol.* 36, 228–239. doi: 10.1111/j.2044-8317.1983.tb01129.x

Westermann, R. (1985). Empirical tests of scale type for individual ratings. *Appl. Psychol. Meas.* 9, 265–274. doi: 10.1177/014662168500900304

Woike, J. K., Hertwig, R., and Gigerenzer, G. (2023). Heterogeneity of rules in Bayesian reasoning: a toolbox analysis. *Cogn. Psychol.* 143:101564. doi: 10.1016/j.cogpsych.2023.101564

Wundt, W. (1896). Grundriss der Psychologie [Outlines of psychology]. *Engelmann*.

Yannakakis, G. N., and Martínez, H. P. (2015). Ratings are overrated! *Front. ICT* 2:13. doi: 10.3389/fict.2015.00013

frontiers | Frontiers in Psychology

TYPE Conceptual Analysis
PUBLISHED 12 September 2024
DOI 10.3389/fpsyg.2024.1390233



OPEN ACCESS

EDITED BY
Jan Ketil Arnulf,
BI Norwegian Business School, Norway

REVIEWED BY
Auke Hunneman,
BI Norwegian Business School, Norway
Geir Smedslund,
The Norwegian Medicines Agency, Norway
Jean Charles Pillet,
TBS Business School, France

*CORRESPONDENCE
Roland Mayrhofer
✉ roland.mayrhofer@ur.de

RECEIVED 22 February 2024
ACCEPTED 19 August 2024
PUBLISHED 12 September 2024

CITATION
Mayrhofer R, Büchner IC and Hevesi J (2024) The quantitative paradigm and the nature of the human mind. The replication crisis as an epistemological crisis of quantitative psychology in view of the ontic nature of the psyche.
*Front. Psychol.* 15:1390233.
doi: 10.3389/fpsyg.2024.1390233



# The quantitative paradigm and the nature of the human mind. The replication crisis as an epistemological crisis of quantitative psychology in view of the ontic nature of the psyche

Roland Mayrhofer*, Isabel C. Büchner and Judit Hevesi

Department of Psychology, University of Regensburg, Regensburg, Germany

Many suggestions for dealing with the so-called replication crisis in psychology revolve around the idea that better and more complex statistical-mathematical tools or stricter procedures are required in order to obtain reliable findings and prevent cheating or publication biases. While these aspects may play an exacerbating role, we interpret the replication crisis primarily as an epistemological crisis in psychology caused by an inadequate fit between the ontic nature of the psyche and the quantitative approach. On the basis of the philosophers of science Karl Popper, Thomas Kuhn, and Imre Lakatos we suggest that the replication crisis is therefore a symptom of a fundamental problem in psychology, but at the same time it is also an opportunity to advance psychology as a science. In a first step, against the background of Popper's Critical Rationalism, the replication crisis is interpreted as an opportunity to eliminate inaccurate theories from the pool of theories and to correct problematic developments. Continuing this line of thought, in an interpretation along the lines of Thomas Kuhn, the replication crisis might signify a model drift or even model crisis, thus possibly heralding a new paradigm in psychology. The reasons for this are located in the structure of academic psychology on the basis of Lakatos's assumption about how sciences operate. Accordingly, one hard core that lies at the very basis of psychology may be found in the assumption that the human psyche can and is to be understood in quantitative terms. For this to be possible, the ontic structure of the psyche, i.e., its very nature, must also in some way be quantitatively constituted. Hence, the replication crisis suggests that the ontic structure of the psyche in some way (also) contains a non-quantitative dimension that can only be grasped incompletely or fragmentarily using quantitative research methods. Fluctuating and inconsistent results in psychology could therefore also be the expression of a mismatch between the ontic level of the object of investigation and the epistemic level of the investigation.

KEYWORDS

**replication crisis, quantitative psychology, human mind, epistemology, ontology**

## 1 Introduction

Is the so-called replication crisis in psychology really a crisis that threatens psychology as an academic discipline in any way? Before answering this question, it is helpful to first outline the broader context. The replication crisis affects not only psychology, the focus of this study, but science as a whole, which is why important fundamental questions of philosophy of science are at stake here. The term "replication crisis" summarizes a number of problems that all revolve around the observation that certain results of scientific research cannot be replicated (for a summery, see Romero, 2019). Beginning in the 2010s, it was first noted for isolated, prominent topics—social priming as well as other findings from social psychology (Harris et al., 2013; Klein et al., 2014) and extrasensory perception (Galak et al., 2012)—then systematically across several areas of psychology that a substantial proportion of published studies, approximately between 23 and 62%, cannot be replicated or can only be replicated to a limited extent (Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015). In other disciplines such as medicine (e.g., Ioannidis, 2005), economics (e.g., Camerer et al., 2016), natural sciences and engineering (e.g., Baker, 2016), it has also been found that only some of the published results can be replicated. Since replication of findings is a cornerstone of scientific methodology and the justification of knowledge, the term "replication *crisis*" was used for the observation that many findings cannot be replicated in order to express the notion that this is a—potential—problem (Romero, 2019).

At the same time, methodological problems have been intensively discussed in psychology since the 2000s, above all questionable research practices, i.e., practices that can be used to achieve significant results, from the exploitation of statistical aspects to make results significant, to non-transparent procedures to veil possible problems and present a found result as unambiguous, to the direct manipulation of data to achieve the desired result (for a summary, see O'Donohue et al., 2022). In psychology, the method—above all a quantitative-experimental approach—is generally predominant and confidence in theories is often greater than in the methods, so that the unexpected outcome of an experiment is often attributed to errors in the method, so that instead of modifying or discarding the theory, attempts are made to change the method so that the result predicted by the theory is achieved (Eronen and Bringmann, 2021; Mayrhofer and Hutmacher, 2020). This fundamental focus on methodology probably led to the replication crisis being viewed primarily as a crisis of methodology, in particular of the statistical methods used, and accordingly the solution would also lie in improved statistical methods. For example, the use of frequentist statistics, especially null hypothesis significance testing, was criticized and the increased use of descriptive (e.g., Trafimow and Marks, 2015) or Baysian statistics (e.g., Colling and Szűcs, 2021), a stronger focus on statistical power (e.g., Anderson and Maxwell, 2017; Shrout and Rodgers, 2018), effect sizes (e.g., Flora, 2020), confidence intervals (Amrhein et al., 2019), equivalence testing (Lakens et al., 2018), or reforming the use of the *p*-value (e.g., Anderson, 2020; Benjamin et al., 2018) were suggested as improvements. In addition, methods such as machine learning (Orrù et al., 2020), meta-analyses (e.g., Sharpe and Poets, 2020), structural equation modeling (e.g., Kline, 2023), multiverse (e.g., Oberauer and Lewandowsky, 2019) or speciation curve analyses (e.g., Steegen et al., 2016) were proposed as methods with which the replication crisis could be countered.

Besides these many proposals relating to statistical aspects—i.e., the way in which data is processed and interpreted numerically and mathematically—a second perspective aims at social-organizational aspects of the scientific process, namely proposals to prevent questionable research practices, to prevent publication bias or the file drawer problem, to mitigate the publish-or-perish problem, or to improve the institutional framework conditions of research in order to counter incentives for fraud (e.g., Asendorpf et al., 2013; Francis, 2012; Greenfield, 2017; Irvine, 2021; Koole and Lakens, 2012; Korbmacher et al., 2023; Lilienfeld, 2017). A third direction is aimed at the theories that underlie research (Fiedler, 2017 and 2018; Lilienfeld and Strother, 2020; Oberauer and Lewandowsky, 2019; Scheel et al., 2021; Scheel, 2022), but the focus there is on the fact that these proposals do not deal with individual specific theories and their content, but argue—on a meta-level, as it were—that generally better theories are needed.

Despite this extensive discussion revolving around the replication crisis and the many suggestions on how to counter the replication crisis, there is no evidence of specific negative institutional-systemic consequences, e.g., no psychological institutes at universities have been closed, and the performance and functioning of academic psychology has not declined either, in the sense that no less output in the form of articles has been produced than before the replication crisis. In fact, there is even evidence that non-replicable studies are cited more often than replicable ones (Serra-Garcia and Gneezy, 2021). From this perspective, then, it appears that the failure to replicate certain findings has had little or even no impact on psychology as an academic discipline. There are also voices that argue that the observation that results cannot be replicated is not problematic at all (Haig, 2022; Maxwell et al., 2015; Schmidt and Oh, 2016; Stroebe and Strack, 2014). Yet this perceived need to defend the status quo and counter ideas of a crisis in itself and, conversely, the many suggestions on how to counter the replication crisis, suggest that there is an important and fundamental issue at stake here. The present study argues, first, that at its core the replication crisis is not a methodological or social-institutional crisis, but rather—following a suggestion by Morawski (2019)—an epistemological crisis revolving around the question of how to justify the knowledge that psychology generates. Second, while what has been called the replication crisis is indeed a substantial problem for psychology, this crisis also opens up the possibility of clarifying fundamental epistemic and ontic questions in psychology. The ontic implications associated with this epistemological crisis are also discussed, i.e., whether the core of the replication crisis—and in a broader sense of psychology as a scientific discipline—is to be found in the very nature of the psyche itself, and whether the research methods used are not or only partially capable of grasping this nature.

Three classics of philosophy of science—Karl Popper, Thomas Kuhn, and Imre Lakatos—provide a promising framework for analyzing the replication crisis from a philosophy of science perspective. Although these theoretical approaches focus on different aspects and are also considered incompatible in some cases, together they can offer explanations that make the replication crisis more comprehensible, as will be shown below. The focus here is primarily on epistemological aspects, and accordingly the replication crisis is viewed here primarily as an epistemological crisis and less as a methodological crisis, more precisely as a consequence of an inadequate approach to the human mind as the object of investigation in psychology. The

replication crisis, as well as many proposals on how to deal with it, are very much focused on quantitative aspects, namely the quality of data and its statistical analysis, but at the same time it remains doubtful whether these proposals have led to improved replicability. Therefore, this study proposes the possibility that the human psyche—possibly due to its very nature—at least partially resists access through a quantitative perspective and approach, of which the replication crisis may be a symptom. Therefore, if the epistemological approach to the psyche through a primarily quantitative perspective does not fit the fundamental ontic structure of the psyche, it is to be expected that the corresponding results are ambiguous and instead point to a fundamental problem, i.e., that an epistemological crisis occurs.

## 2 Perspectives from philosophy of science and their consequences for psychology

### 2.1 The Popperian perspective: failed replication as the opportunity to improve theories

According to Popper's (1959/2005) Critical Rationalism, a perspective on science which is widespread in academic psychology, the failure to replicate certain findings is part of the "normal" and even desirable progress in science (see also Derksen, 2019; Laws, 2016; Keuth, 2005; Rowbottom, 2011). Falsified hypotheses are rejected and hypotheses that have withstood attempts at falsification are retained—at least for the time being, and at least according to Popper's idea of ideal science. Many proposals concerning the replication crisis accept the basic epistemological premise, largely based on Fisher's (1935/1974) and Popper's (1959/2005) influential books, which have substantially shaped the methodology and the self-conception of psychology, that reproducibility is one of the basic requirements of science in order for its results to be justifiably considered knowledge. Popper started from the so-called problem of induction, i.e., the question of whether and how inductive conclusions can be justified. On the one hand, a large number of similar observations allow the prediction that the same phenomenon will also be repeated in the future, but on the other hand, recourse to past observations cannot guarantee that this will also apply to the future. Popper "solved" the problem of induction—a more detailed analysis of this intricate problem lies outside the scope of this article (see, e.g., Agassi, 2014; Musgrave, 1993; Swann, 1988)—by reversing the problem, so to speak, and postulating instead that theories should not be verified but rather falsified. Therefore, replications, which in principle are the repetition of an observation, play an epistemologically subordinate role because they "only" confirm, i.e., "verify," previous observations, and according to Popper verification is impossible in principle. Verifications do support theories, and theories that are supported by many observations—or, according to another interpretation, that have withstood many attempts to disprove them—are considered more likely to be true, but theories cannot be proven by repeated identical observations, only be disproved by conflicting observations.

This raises the question of how to interpret a replication: Is it an attempt at verification that adds another confirming observation to a theory if the replication is successful, thus increasing its probability of being true? And if so, how many successful replications are necessary for a theory to be accepted as true with some probability? In other words, can knowledge be quantitatively justified? Conversely, is an unsuccessful replication attempt—perhaps even a single unsuccessful replication—to be equated with a refutation of the theory in question? Or is an unsuccessful replication merely the lack of confirmation of a theory that, according to Popper, has a lower epistemological value than a direct refutation? Although the answer may depend on the specific theory in question, clarifying these questions is crucial to understanding the replication crisis and its epistemological dimensions.

It is also necessary to clarify what exactly is meant by replication. In psychology, people—and not inanimate matter—are usually studied, and therefore a completely exact replication is impossible in almost all cases because study participants are changed by their very participation, so that a study cannot be conducted with the same people and new participants necessarily differ from the previous ones. Epistemologically, it could be argued that people often differ only slightly, at least in a particular aspect which is of interest in a study, that said aspect is distributed in a certain way, which allows a statistical approach, or that with a sufficiently large sample the mean can be used as an estimator, and it is therefore justified to speak of replication as long as the study design itself remains unchanged. Interestingly, all of these points contain a more or less clear quantitative component: This is evident in statistical aspects, but statements about the size of differences also imply at least a rudimentary quantitative understanding. This is a first indication that the human mind—at least in certain aspects—is regarded as quantitatively constituted in psychology and thus meaningfully accessible to quantitative methods.

However, even if one accepts these arguments concerning replication, the question arises as to the time periods for which such equality is assumed, as cultures and societies, and therefore also people, change over time—and change to such an extent that psychological processes may also be affected (e.g., Hutmacher and Mayrhofer, 2023). This problem obviously exists with standardization and calibration, for example with intelligence or personality tests that have to be updated over time, or with test–retest reliability in general, so that the question arises as to whether other psychological processes—e.g., cognition, motivation, or emotion—also change over time. On a more practical level, exact replications also appear difficult, as they may be carried out by other investigators, in translation, with different materials, or in other cultures, all of which may influence the outcome. This is illustrated by the well-known WEIRD bias in psychology, according to which the majority of the results of psychological research are obtained from a very specific group, namely American undergraduates, that is hardly representative of humanity as a whole, but the results are often regarded as universally valid (Henrich et al., 2010). Thus, from how much deviation do we no longer speak of replication? Even this brief sketch shows that the question of the basic conditions for replication is not easy to answer.

From a different perspective, however, another problem can be identified here that is even more fundamental in terms of epistemology. If replications are suitable for supporting or refuting the validity of theories, then this presupposes that the way in which the associated empirical observations are carried out and measured is also suitable for answering the theory or research question in a meaningful way. Otherwise, neither corroboration nor refutation would be possible, because the measurements, data, and results as well as the conclusions drawn from them would have no meaning then and could not be interpreted as corroboration or refutation either.

Now, all studies that were examined and replicated for the original replication project (Open Science Collaboration, 2015) and its continuation (Camerer et al., 2018) were experimental psychological studies in which a quantitative methodology was used. This fact in itself is remarkable, because these experiments were intended to be representative of (experimental) psychology (Open Science Collaboration, 2015) or they appeared in prestigious journals (Camerer et al., 2018). Furthermore, the experiments were also chosen for practical reasons, namely, that "[t]he key result had to be represented as a single statistical inference test or an effect size" (Open Science Collaboration, 2015 p. 2) or that there was a "clear hypothesis with a statistically significant finding" (Camerer et al., 2018, p. 1). The analysis of the replications carried out and the subsequent interpretation that many previous findings could not be replicated was also quantitative. Since it is difficult to specify clear quantitative criteria for when a replication is successful or not (e.g., Chambers, 2017; Cumming 2008; Gelman and Stern, 2006; Open Science Collaboration, 2012, 2015; Simonsohn, 2015; Verhagen and Wagenmakers, 2014), the problem described above of how replications are to be classified in terms of epistemology theory is further exacerbated.

Although it remains to be discussed whether an unsuccessful replication represents a refutation, the failure to replicate findings is critical in Popper's view as the theories in question are not corroborated and thus prone to rejection and elimination from our pool of theories, being replaced by theories that are better supported by repeated observations. From this perspective, the replication crisis is not a crisis at all but rather a process that increases our knowledge by demonstrating that certain theories are false or at least cannot be corroborated by repeated observations, increasing their probability of being false. Therefore, notwithstanding the many problems of the various forms of Critical Rationalism (e.g., Agassi, 2014; Keuth, 2005; Rowbottom, 2011), the Popperian perspective offers a different view on the replication crisis: From this point of view, the replication crisis can be seen as a corrective pruning process because it allows the discovery of potentially false theories, which can be removed from our pool of theories, thus creating space for new theories that are closer to truth.

## 2.2 The Kuhnian perspective: unexpected observations as a harbinger of a model crisis

Karl Popper and Critical Rationalism assume that there is an objective truth and, based on this, that knowledge is also objective. In contrast, Kuhn (1966; see also Marcum, 2005; Nickles, 2003) strongly emphasizes the social dimension of science as a collective process. In a nutshell, Kuhn assumes that science is not a more or less linear process in which we get steadily closer to truth over time. Instead, a cyclical model is postulated in which different paradigms[1] replace each other. Once a paradigm has established itself and is considered to be true, further research then takes place within this paradigm—the so-called "normal science." This is not only a purely "rational" process, in which exclusively only aspects that are directly related to the object of knowledge are decisive, but also other, mainly social, factors play a role in which this is not the case and which instead indirectly affect a paradigm, e.g., influential persons who control the flow of resources or the allocation of academic positions and who can therefore influence other researchers, or general cultural and social conditions that favor thinking in a certain direction and marginalize other directions. However, at some point the first observations are made that do not appear compatible with the prevailing paradigm—the first signs of a so-called "model drift." Initially, these observations are simply ignored or labeled as anomalies, but over time there is mounting evidence that the prevailing paradigm does not represent the (whole) truth—what is called "model crisis." Eventually, the prevailing paradigm can no longer be maintained and a "model revolution" occurs in which a new paradigm prevails, which then becomes the new normal science. In this process, it must be taken into account that not only "rational" factors directly related to the object of knowledge play a role, but also—as already mentioned—social or cultural factors, such as when influential persons who upheld a paradigm no longer (can) perform this function.

According to Kuhn's model, which is less epistemologically and more sociologically oriented, crises that give rise to doubts about previous knowledge are processes that occur regularly and more or less systematically. From a formal point of view, i.e., if the cycle described above is regarded as a theory that can describe and predict the course of science, it may be assumed that the replication crisis could signify a model drift or even a model crisis as unexpected observations have emerged.

These observations are unexpected because, according to the current state of knowledge—i.e., high-ranking published studies in which a specific psychical[2] phenomenon is described—it should be assumed that this knowledge is reliable and can therefore be replicated by and large. There are three possible reasons why this is not or only partially the case: first, the original knowledge, i.e., the original studies, is false, so the failed replications are correct. Second, the original studies describe true phenomena and theories but the replications are—for whatever reason—untrue. These two possibilities could presumably be clarified by carrying out many replications, perhaps also with additional variations, in order to be able to determine the influence of different effects and variants (e.g., Breznau et al., 2022; Muñoz and Young, 2018; Silberzahn et al., 2018; Simonsohn et al., 2020; Steegen et al., 2016; Young, 2018). If there are clear tendencies, it would be possible to recognize whether the effect or mechanism postulated in the original study actually exists in a general form or whether it is merely an individual situation that resulted from certain idiosyncrasies. Therefore, these possibilities can be dealt with within the currently prevailing paradigm, i.e., the so-called normal science.

1 The terms "paradigm" and "model" are usually employed interchangeably. However, the phases in Kuhn's model are commonly referred to as "pre-science," "normal science," "model drift," "model crisis," "model revolution," and "paradigm shift," with "*paradigm* shift" being used instead of "*model* shift."

2 As suggested by Uher (2021), the term "psychical" is used here as adjective for phenomena that relate to the psyche itself, e.g., motivational, cognitive or emotional mechanisms. In contrast, "psychological" is used for research into psychical phenomena, i.e., experiments and other studies or theories on, e.g., motivational or emotional phenomena.

A third possibility, however, is that it is not possible to say with any certainty whether the original study or the replication is true. This possibility can be attributed to the assumption—as explained later in this study—that both the original studies and replications may not be suitable for adequately grasping the psychical phenomenon of interest. Such an inadequate fit between phenomenon and research method leads most likely to inexplicable results in the observation and analysis of the phenomenon, which cannot be understood within the paradigm of normal science because the theoretical and conceptual foundations are not sufficient. This connection was demonstrated by Kuhn (1966) and Feyerabend (1975), primarily using examples from astronomy, and even if the controversial question of whether a general theory of how science works can be derived from this is excluded (e.g., Farrell, 2003; Oberheim, 2006; Preston, 1997), these cases illustrate the possibility of a model crisis and a paradigm shift.

For psychology and the replication crisis, it is now relevant that the methods used reflect the paradigm within which they are used. Therefore, unexplained results may indicate that the interplay of basic theoretical assumptions and methods is not appropriate to the phenomenon under investigation, casting doubt on the underlying paradigm, thus possibly heralding a model crisis or even model shift in psychology.

So, while Kuhn's theory can explain the systemic and social reasons why a paradigm shift occurs in science, it does not, in terms of the specific scientific content, provide explanations as to why the "anomalies" challenge the prevailing paradigm. While this complex fundamental question (e.g., Fuller, 2003; Lakatos and Musgrave, 1970; Toulmin, 1972) lies outside the scope of this analysis, the model of paradigm shifts nevertheless seems to imply that some theories somehow fit empirical observations better than others. Abstractly speaking, Kuhn's model thus always contains an epistemological crisis, and since—as shown above—the replication crisis is an epistemological crisis, it can consequently be interpreted in Kuhnian terms as a model crisis or even model drift. Furthermore, merging the more specific epistemological level, as described above in Popper's model, with Kuhn's model, justification of knowledge plays an important role in both cases because, epistemologically, failed replications can lead to an undermining of existing knowledge, which in turn anticipates a model crisis and, eventually, a model revolution and paradigm shift.

Furthermore, Kuhn's model may be supplemented by the observation that over time models and procedures can lose their connection to the actual object of investigation and instead only revolve around themselves (Elster, 2016), meaning that in the last phases before a paradigm shift, the traditional way of doing science—"normal science," in Kuhnian terms—loses its vitality and fossilizes. Interestingly, when this happens, there can also be a tendency toward "mathematical sophistry," so that the methodological tools also lose their relation to the phenomena being investigated and instead become a purposeless "toy" (Elster, 2016, p. 2182).

## 2.3 The Lakatosian perspective: the role of methodology in psychology

Lakatos's (1978; see also Larvor, 1998) philosophy of science focuses on the concept of the so-called "research program." This is a central set—called the "hard core"—of related, interdependent axioms, concepts, theories, and possibly also methodologies, which provide the foundations, guidelines, and directions for research and that cannot be abandoned or altered without compromising the research program itself. Around the hard core, there is a protective belt of so-called auxiliary hypotheses, which usually concern methodological aspects and deal with anomalies or observations contradicting or inconsistent with the central assumptions of the hard core. Rather than disputing the hard core itself, which would challenge the very foundations of the research program, problems that arise from such conflicting observations—in Kuhnian terms, the "anomalies"—are rerouted to the protective belt. Thus, instead of modifying or abandoning the central assumptions of the hard core, attempts are first made to defuse "problematic" observations by dealing with them at the level of auxiliary hypotheses, i.e., usually at the methodological level, trying to explain said observations by methodological errors, inaccuracies, or other shortcomings. If this is not or no longer possible, the auxiliary hypotheses can be modified so that "problematic" observations can be explained without compromising the hard core.

There are, however, two crucial points: First, the auxiliary hypotheses and the protective belt must somehow be conceptually related to the hard core, i.e., the auxiliary hypotheses and the protective belt must not be incompatible with the hard core because otherwise they could not protect the hard core at all but would rather challenge it. Second, the line between fundamental concepts and the hard core and auxiliary hypotheses and the protective belt is not always clear-cut. This makes it difficult to decide if modifications affect only the auxiliary hypotheses, i.e., if the protective belt functions actually as protection of the hard core or if the ramifications are so far-reaching and profound, going beyond the protective belt, that the hard core itself is affected by assumptions that were originally meant to protect it. Accordingly, the hard core is only abandoned if conflicting data and contradictions can no longer be rerouted to and resolved within the protective belt.

Complicating matters further, the extent of a hard core is a matter of discussion. In the case of psychology, there is no clear hard core as focal point for the whole discipline or its branches because the subject matter, namely human mind and behavior, is very vast and diverse and there is presently no fundamental or all-encompassing theory which might provide a coherent framework for a research program in the Lakatosian sense. For much of the 20th century, behaviorism can be regarded as research program because the fundamental idea that virtually all behavior can be explained in terms of stimulus, response, and contingencies provides a coherent and all-encompassing theory as the basis for a research program. Evolutionary psychology and behavioral neuroscience may be regarded as attempts to establish a hard core in the Lakatosian sense for psychology, because both operate from the basis of a single fundamental theory, namely that mind and behavior can be explained by evolutionary or biological processes, respectively. However, none of these approaches has gained near-universal acceptance or has produced decisive results to dominate academic psychology.

On a less global level, certain paradigms could be seen as research programs, such as the idea in neuropsychology that certain behaviors, personality traits, or mental disorders can be localized in certain places in the brain (e.g., Corr et al., 2013; Dolan and Park, 2002; Shenal et al., 2003; Schretlen et al., 2010). In cognitive psychology, the testing effect can be interpreted as a research program because, built on a fundamental assumption, namely the effect of retrieval, further theories are grouped together (e.g., Rowland, 2014; Schwieren et al., 2017) which—and this is the crucial

point—would immediately lose their validity if the effect of retrieval as a common focal point would turn out to be false.

Despite the lack of a hard core of fundamental and universal theories in contemporary psychology, there nevertheless seems to be some kind of unifying factor which provides coherence to psychology as an academic discipline, namely the focus on a methodology that is characterized by experimental, quantitative, and empirical approaches (Mayrhofer and Hutmacher, 2020). This observation is crucial for any analysis in Lakatosian terms because it can be argued, on the one hand, that the dominance of this methodology constitutes a research program by providing a coherent frame within which research in psychology is conducted. On the other hand, the hard core of a Lakatosian research program is not—at least not primarily—characterized by a certain methodology *per se* but rather by central concepts and theories, and the preferred or characteristic methodology reflects the supposedly best way to investigate the central concepts and theories.

Therefore, it seems that the quantitative-experimental methodology fulfills a dual role: First, it acts as a "protective" belt of auxiliary hypotheses that virtually defines how psychical phenomena are approached, thus shielding the core from questions or problems which cannot be approached quantitatively, empirically, or—to a lesser extent—experimentally. Consequently, psychological phenomena that are not accessible to such a quantitative-experimental approach are sidelined and eclipsed by the vast research conducted according to those very principles. Second, at the same time, there is no fundamental universal theory that could explain all these phenomena and thus serve as the focal point and hard core of a research program. Since such a blank space cannot hold together a research program, methodology takes on this task as a substitute, as it were. Taken to its logical conclusion, this means that the methodology protects itself—which is a somewhat paradoxical statement that will be explained in more detail below.

However, while it remains unclear what the hard core actually is, the shielding function of the protective belt can also be analyzed from the question of whether a research program is—in Lakatosian terms—progressive or degenerative (Lakatos, 1978). Modifications in the auxiliary hypotheses can prompt further advancements and refinements within the research program, thus strengthening the hard core and the fundamental theories by clarifying problems or correcting minor errors and defects in the central concepts and theories. In this case, the research program is considered progressive because it produces new knowledge and its explanatory power is increased. If, by contrast, modifications in the auxiliary hypotheses do not improve the hard core but simply serve to shield it from conflicting observations, thus actually decreasing the scope and explanatory power of the fundamental theory, the research program is considered degenerative.

Lakatos (1978) discussed the relationship between methodology and the hard core of the fundamental theories in terms of the so-called positive and negative heuristic. Based on a more differentiated interpretation of modus tollens than in Critical Rationalism, the negative heuristic states that observations inconsistent with the fundamental theories should not be immediately regarded as falsifications, thus protecting the hard core. The consequence is that discussions about how challenging observations should be interpreted and handled often take place at the level of the auxiliary hypotheses, i.e., in the protective belt, which comprises the methodology as well. The positive heuristic, on the other hand, acts as a methodological framework within which research is carried out. It provides certain strategies, tools, and techniques to solve problems and answer questions that are typical for the research program. Successful approaches yielding fruitful results usually become the methods of choice precisely because they have shown their efficacy and thus promise to be able to answer further questions as well. As a consequence, however, relying on a "tested" and "safe" methodology also implies or even determines what kind of problems and questions are addressed—namely those compatible with the preferred methodology.

Against the background of Lakatos' theory, the replication crisis can be interpreted as follows. According to Lakatos, if a substantial number of findings cannot be replicated—i.e., anomalies occur, in Kuhnian terminology—this problem is first dealt with at the level of the protective belt. This assumption fits with the observation that the discussion on the replication crisis primarily revolves around the level of methods, i.e., improving data quality and analysis. This discussion takes place on the level of the protective belt, because being about methodology it is about access to psychical phenomena and not about the psychical phenomena themselves. Therefore, this discussion reflects a fundamental epistemological problem, namely the question of how to gain appropriate access to psychical phenomena, i.e., the object of investigation in psychology.

However, since—as explained above—it remains unclear and vague what exactly the hard core of psychology consists of and instead the methodology, i.e., a quantitative approach, vicariously assumes the role of giving the discipline a structure and the research activities a direction in the sense of a Lakatosian research program. However, if the methodology of psychology is called into question, it is not only the protective belt that is affected, but also the very core. Due to this peculiarity, fractures in the protective belt thus also affect the core of psychology, and these potentially far-reaching consequences point to a model crisis in the Kuhnian sense.

## 2.4 The quantitative paradigm and the replication crisis

The questions of whether a research program—in Lakatos' sense—is progressive or degenerative, and whether a positive or negative heuristic is present, can be applied to the replication crisis. Many suggestions on how to counter the replication crisis revolve around the improvement of statistical methods, i.e., quantitative methods. Against the background outlined above, this is important in several respects:

First, this discussion can be interpreted as a typical methodological discussion that takes place at the level of the auxiliary hypotheses, precisely because the methods of a research program are the focus and not the underlying theories of psychical phenomena themselves. Second, the discussion about means to solve the problems raised by the replication crisis is characterized by ambivalence: On the one hand, if these proposals are successful, these changes in methodology, i.e., at the level of the auxiliary hypotheses, would improve the ability of the hard core to deal with problematic observations, which would be progressive. On the other hand, it is doubtful whether the elimination of a problem—lack of replicability—can actually be seen as generating new knowledge and increasing the explanatory power of the theories of the hard core. From this perspective, it would therefore be more appropriate to speak of a defensive discussion that attempts to solve problems by eliminating anomalies, which would qualify the research program as degenerative.

Third, this is all complicated by the fact that it is unclear what the hard core actually is and what its basic assumptions and theories are. However, if a large part of the discussion on how to counter the replication crisis revolves around methodological questions, and if these methodological questions are discussed independently of the content of psychical theories, the auxiliary hypotheses in the protective belt do not protect the hard core of psychical theories but rather the methodology itself. Improving the methodology without tying it to genuine psychical theories is epistemologically problematic because then the methodology revolves around itself and the research program becomes degenerative.

Viewed more generally from a philosophy of science perspective, a mismatch between methodology and psychical theories can also be interpreted as an insufficient or inadequate understanding of the ontic nature of the object of investigation—in this case the psyche—from which a set of fundamental interrelated epistemic problems arises. Although the object of study in psychology is obviously the psyche, a precise definition of this term is difficult and controversial (e.g., Mayrhofer and Hutmacher, 2020). This difficulty in finding a common denominator for cognitive, emotional, motivational phenomena and the like is a first indication that a fundamental issue is at stake here. For the purposes of this study, however, it is sufficient to understand "psyche" unspecifically—and somewhat tautologically—as the totality of psychical phenomena as studied by psychology. The ontic nature of the psyche refers to the fundamental being or essence—in a philosophical sense—of the psyche itself and not how it functions. Classical concrete ontic questions, such as the conditions of the possibility of being (here: of the psyche) in the abstract sense but also the mind–body problem (Weir, 2024) or questions about the nature of consciousness (Rowlands, 2001) or emotions (Soteriou, 2018) can be largely excluded here, because the focus is on the abstract relationship to the epistemic level.

The aim of ontology (e.g., Effingham, 2013) is not only to understand the nature of being and what it means for something to exist (in a certain form), but also to categorize (ontic) entities, to clarify their relationship to each other and the principles governing their functioning. By addressing the most fundamental ontic aspects of an object (of investigation), ontology also provides a frame of reference for other disciplines by clarifying the fundamental structures and conditions that constitute the object of investigation. Epistemology deals, in short, with everything that has to do with the nature of knowledge, its generation and justification (e.g., Carter and Littlejohn, 2021). What we know and can know about an object is therefore not only an epistemic question—e.g., which methods can be used to approach the object, to what extent the object is recognizable at all, or how the object can function in principle—because the answers to these questions are obviously (also) enabled, determined and limited by the ontic nature of the object. Thus, the ontic structure of an object necessarily affects our epistemic understanding of it and knowledge results if the ontic and epistemic levels are in agreement (Bachelard, 1974; Sandkühler, 1991). For the way in which such an object is constituted in terms of its ontic structure also determines the possibilities of grasping it epistemically. One of the reasons why such an investigation is possible is that the investigating entity, i.e., humans, must somehow—and the exact nature of this relationship is disputed—be compatible with the object of investigation due to its own ontic constitution, because otherwise the investigating entity would have no way of understanding the object of investigation. The ontic relationship between the object of investigation and the investigating entity thus determines the epistemic possibilities of the investigating entity to grasp and understand the object of investigation (for a summary, e.g., Jacquette, 2014; Morawski, 2019; Steup, 1996; Steup and Ram, 2024).

However, if the epistemic and ontic levels are mismatched far-reaching and serious problems can arise, for example if assumptions are made on the epistemic level about how to approach the object of investigation that do not match the ontic structure of the object of investigation, are incompatible with it, or even contradict it. First, the object of investigation and how things work cannot be understood, or can only be understood inadequately, or in a distorted way. Second, as a direct consequence, the unreliable knowledge thus produced and obtained is not suitable as a basis for making correct predictions, interventions, and manipulations, as this knowledge reflects reality only inadequately, distortedly, or even falsely. Thus, the mismatch between the knowledge produced and experiences in reality becomes evident. Third, this results in problems in justifying the knowledge produced in this way—even if it is partially correct and reliable—because it is not systematically correct, but at best selective and possibly for unclear, random reasons. This means, fourth, that a scientific discipline is thus likely to produce anomalies and enter into a crisis (in Kuhnian terms) or to stagnate or degenerate as a research program (in Lakatosian terms).

If the replication crisis, as argued above, is indicative of a fundamental epistemic problem in psychology, this problem could lie precisely in such a mismatch between the epistemic and ontic levels. In concrete terms, this means that a fundamental aspect or dimension of the ontic nature of the psyche may not be understood, insufficiently understood, or misunderstood and thus neglected or inadequately addressed in research. As this dimension is not considered in research, but— presumably— is nevertheless present and affects the functioning of the psyche, research and its results are influenced by this unknown and unconsidered factor, which in turn could explain the anomalies and fluctuating results seen in the replication crisis. In other words, the replication crisis may be interpreted as an epistemological crisis rooted in an inadequate understanding of the ontic constitution of the psyche, leading to a mismatch between methodology and the epistemic level on the one hand, and the nature of the psyche as an object of inquiry on the other.

## 2.5 Psychology and the nature of the human mind

Considering the highly quantitative nature of psychology as a whole, as well as the proposed solutions to the replication crisis, which very often focus on quantitative aspects, this could be an indication that the root of this mismatch lies precisely here. This means that the human psyche might not be or only partially be accessible to investigation by quantitative methods—or theories based on quantitative thinking—due to its very ontic constitution. As a consequence, improvements in quantitative methods cannot resolve or mitigate the problems of the replication crisis.

That the replication crisis is a symptom of a fundamental problem in psychology, and that it revolves primarily around a methodology that is by its nature primarily quantitative, thus suggests that the mismatch between the ontic and epistemic levels

may be rooted precisely in the quantitative nature of the methodology. This is because adequate access to the object of investigation using quantitative methods presupposes that it can also be grasped quantitatively. If problems arise, it is possible that the object of investigation cannot be grasped quantitatively because its ontic structure is such that certain aspects somehow elude such quantitative access. This suggests that the psyche contains a non-quantitative dimension, meaning the following: Ontological categories are an extensive and complex fundamental topic of philosophy on which there is little agreement (Perović, 2024; Westerhoff, 2005). Although quantity—i.e., how many?—has been considered a fundamental ontological category since Aristotle, what matters here is not what quantity the psyche—or its subsystems and mechanisms—has, but rather that it is quantitatively accessible at all. In order to be quantitatively accessible, the psyche must possess the ontic property of quantitativeness—to be quantitative—that is, to be composed and accessible in quantitative form and to be expressible and conceivable in quantitative, numerical terms. This does not mean that (latent) constructs such as intelligence or certain personality traits are represented in quantitative-numerical terms—and the difficulties in this endeavor are possibly another indication that the psyche contains a non-quantitative dimension—because this is merely an attempt to grasp something quantitatively at the epistemic level. And this attempt does not necessarily guarantee that intelligence or personality traits—apart from their controversial ontic status anyway—actually *are* quantitative in their ontic nature *eo ipso*. The same applies to attempts to grasp and understand subjective experience, aesthetic perception, dreams, unconscious processes and the like through psychological research. This problem is further exacerbated by the fact that there is no consensus on what the nature of the psyche actually is, as illustrated by the multitude of different ideas ranging from Plato's concept of a tripartite soul to current neuroscientific concepts. Interestingly, these concepts do not take into account the question of a possible quantitative dimension of the psyche. For concepts prior to, say, the 19th century—i.e., more or less the beginning of psychology as a science in the modern sense—this is hardly surprising, since, generally speaking, until that time there was little or at least much less thinking in quantitative terms. However, for more modern concepts, which are based more on thinking in quantitative terms, as is typical of modern science, it is quite surprising if such a fundamental question was or is not explicitly discussed, but rather—more or less implicitly—assumed. Although modern ideas of the psyche, such as in psychometrics, behavioral economics, or neuroscience, work with quantitative methods, there has hardly been any discussion to date as to whether this also implies that the psyche is also—in whatever form—quantitatively constituted.

The question of how such a possible non-quantitative dimension of the psyche is to be understood lies beyond the scope of the present study for two reasons: First, answering this question requires extensive research, and second, the aim of this study is to explore quantitativeness as a possible ontic category of the psyche from a philosophy of science perspective and to elaborate the implications for psychology as a scientific discipline. Quantitativeness as a possible ontic category of the psyche, and in particular the property of "non-quantitative" as an explanation for difficulties such as those made visible by the replication crisis, is therefore primarily a matter of identifying a fundamental philosophy of science problem of psychology as a scientific discipline and making it recognizable as a problem. A more precise definition of this problem, describing its specific characteristics and then developing possible solutions are steps that necessarily follow.

Thus, this study raises the possibility that the ontic structure of the human psyche contains a dimension that is not quantitatively constituted and therefore to a certain extent eludes quantitative access. This does not mean that a phenomenon such as intelligence or a cognitive mechanism cannot be approached quantitatively in some form—in the case of intelligence this actually works quite well—but there is always the possibility that decisive aspects are not covered, which can lead to inexplicable variance, as exemplified in the replication crisis. In other words, it is possible that an epistemological crisis can be traced back to an insufficient epistemic fit with the underlying ontic structure, which possibly contains a non-quantitative dimension that could explain that insufficient fit. The nature or ontic structure of this something—be it directly intelligence or personality itself or a currently unknown underlying phenomenon—is relevant in this context, since it is the ontic structure that provides the basis for the phenomenon to be epistemically accessible and comprehensible. The same applies to cognitive, motivational, or emotional mechanisms as well as to consciousness, all of which can be observed—as surface phenomena, so to speak—but whose ontic structure is still completely unclear.

Three examples can be used to illustrate, at least to some extent, what such a non-quantitative dimension might look like: First, questions about qualia (e.g., Nagel, 1974; Tye, 2021) or meaning (e.g., Flanagan, 2007), which are fundamental to human psychical experience, have so far eluded not only any quantitative approach, but also a precise determination of their ontic nature. Second, the same applies to language, which in principle cannot be quantified either, because it works with meanings (e.g., Lycan, 2019; Platts, 1997). Third, Jaeger et al. (2024) have argued that agency, cognition, and consciousness cannot be computational or formalized or captured by algorithmic approaches. These examples thus suggest that a non-quantitative dimension exists in the ontic structure of the human psyche, even if it cannot yet be described in more detail.

The question of the ontic structure and nature are closely related to another—unsolved!—fundamental ontic problem of psychology, namely the mind–body-problem. Quantitativeness as an ontological category and the assumption of a non-quantitative dimension of the human psyche is in principle compatible with all three fundamental positions: In idealistic positions, a non-quantitative dimension must be thought of as immaterial, which in turn raises the question of what this looks like in concrete terms. With materialistic positions, the additional question arises as to how a non-quantitative—or quantitative, for that matter—dimension can be derived from a material basis. Dualistic positions are faced with the problem of which side—or possibly both?—quantitativeness is associated with, whether it manifests itself differently in each case, and what the interaction looks like in concrete terms.

## 3 Discussion

Mathematics is magic, literally and metaphorically. Literally, because magic attempts to depict the world in some form using abstract symbols and to change that what they represent by manipulating these symbols. In mathematics, concrete things or

relationships are also represented in abstract form, namely by numbers and mathematical operations, and the manipulation of this representation makes it possible to make actual changes in the world—and this very often works. And is it not, metaphorically speaking, "magical"—in the sense of astonishing, because this connection is currently neither ontically nor epistemically fully explicable (e.g., Crump, 1992; Horsten, 2023; Shapiro, 2000)—that complex facts of the concrete, material world can be expressed, via universal laws, in abstract and seemingly unambiguous form as numbers and that the manipulation of these numbers can in turn influence the material world?

Against the background that the ontic status of numbers and mathematical operations is still as unclear as their epistemic possibilities and limits, the question arises in a discipline such as psychology, which relies very heavily on quantitative methods, whether there are limits to the use of quantitative methods, where these limits might lie, and what this means for psychology in general as an academic discipline.

Before discussing the implications of this assumption below, it should be noted that the present study is not intended to be prescriptive and no statements are made here about how psychology or, more generally, science should operate. Such claims, as advocated by Critical Rationalism or Logical Empiricism, are now regarded as outdated by philosophy of science and inappropriate for a complex endeavor such as science (e.g., Bird, 2013). Instead, the aim of this study is to identify and discuss possibilities concerning a fundamental problem, i.e., to explore what aspects that have been less or not yet addressed could also be relevant for psychology. Furthermore, it should be noted that science, and thus also psychology, is extremely complex, so that considerations of a general nature, such as those made here, necessarily only represent a rough and abstract outline.

The question of whether the human psyche is non-quantitative or contains a non-quantitative dimension in addition to a quantitative dimension is obviously extremely complex and extensive and goes far beyond the scope of the present study. Moreover, the term "non-quantitative" initially only represents a negative demarcation and an antithesis to the idea that the psyche is exclusively or primarily quantitative. The term "non-quantitative" is not intended at this point to provide a more detailed definition of what such a non-quantitative dimension might look like in concrete terms. On the one hand, this would have to be the subject of a comprehensive discussion from the perspective of various disciplines, which obviously goes far beyond a single study. On the other hand, it is equally unclear what is actually meant by "quantitative"—as a quantitative dimension of the psyche—and what it might actually look like if the psyche functions in a quantitative way. Approaching and possibly clarifying this problem would not only shed light on a fundamental question, but would also put psychology as a discipline on a better footing, as it can be assumed that such knowledge would also change our understanding of how psychical mechanisms work.

If the assumption that the human psyche contains a non-quantitative dimension is correct, then the replication crisis is not an "accident at work" that happened "just like that" due to unique circumstances. Instead, again speaking with Kuhn and Lakatos, such crises must (almost) inevitably occur for systemic reasons, because the object of investigation, i.e., the human psyche, eludes access to a greater or lesser degree due to the methodology used. This lack of fit between an investigated psychical phenomenon and the method used to investigate it in turn means that unexplained factors exert an influence and thus an explanatory gap exists that cannot be closed by normal science, to use Kuhn's terminology.

So, if this interpretation of the replication crisis is correct, there are two reciprocal possibilities for the future: First, if the non-quantitative dimension of the human psyche continues to be (largely) neglected, the replication crisis will continue or repeat itself in a similar form because the or at least one of its root causes has not been addressed. Second, if the non-quantitative dimension of the human psyche is considered more intensively, the replication crisis will be mitigated or will not recur in this form, precisely because the or at least one of its root causes has been sufficiently addressed.

The replication crisis could therefore be a symptom that psychology systematically neglects certain basic ontic conditions of its object of investigation, i.e., the human psyche, or only considers them inadequately. And according to Kuhn and Lakatos, such fundamental problems usually lead to profound changes in a scientific discipline, meaning that it is possible that the replication crisis represents the initial stage of such a model crisis.

The arguments discussed in the present study, which, starting from the epistemological status of replications, lead to fundamental philosophical questions, showing that the replication crisis offers an opportunity to ask fundamental questions about the nature of the psyche. In this sense, the replication crisis is not only a problem that challenges the functioning of the discipline but also an opportunity to clarify the foundations of the discipline and to advance the discipline as a whole by improving its access to the human psyche as its object of study.

Karl Popper, Thomas Kuhn and Imre Lakatos, three classics of the philosophy of science, were used to interpret the replication crisis. Finally, a fourth important philosopher of science, Paul Feyerabend, can be used to illustrate another fundamental aspect: The key message in *Against Method* (Feyerabend, 1975) is that the limiting of methodological approaches restricts access to phenomena and thus hinders scientific progress. According to Feyerabend, methodological approaches and frameworks are not only justified by "rational" reasons but reflect a more comprehensive understanding of the world. Ancient Babylonian science, for example, forms a system that is only partially understandable today because it was embedded in a completely different world view. The same applies to Aristotelian science, whose basic assumptions differ fundamentally from today's science. According to Feyerabend, there are no objective criteria that can rationally justify the superiority of one of these systems. This assumption may or may not be true, but it demonstrates the need to reflect on the general foundations on which science is based because, as the replication crisis suggests, they determine to a large extent how a discipline functions.

However, the results of this study for psychology as a discipline show a peculiarity that has so far received little attention in philosophy and history of science: The falsification of theories, a model crisis, or the degeneration of a research program usually take place at the local level of theories and their concrete content, which relate to specific phenomena. In contrast, this study argues that a very global aspect such as a quantitatively dominated method can be explained by the same mechanisms and can lead to the same situations. It may therefore be that

psychology is a special case that differs significantly from other disciplines. It is possible, for example, that all psychological theories that could not be supported by replications are correct in terms of content but that they are not (fully) accessible with a quantitative methodology. Psychology thus represents an interesting case for the history and theory of science, the further investigation of which could not only advance psychology as a discipline but also provide new insights for the history and theory of science.

Returning to psychology itself and the human psyche, the final question that remains is what the above means in concrete terms for psychology as an academic discipline: There are various suggestions as to how psychology could increase its explanatory power by expanding its range of methods (e.g., Hutmacher and Mayrhofer, 2023; Malich and Rehmann-Sutter, 2022; Wiggins and Christopherson, 2019; Juarrero, 2000). This fits in with Feyerabend's (1975) call not to let the method dominate the research. At the same time, however, the question arises as to whether the possible existence of a non-quantitative dimension in the human psyche does require a different kind of theory that takes this circumstance (better) into account, even if it is not possible to say in advance what this kind of theory should look like.

This study thus suggests that it may be necessary to fundamentally rethink and expand the current framework within which much of psychology operates in order to reflect the full richness of human experience—or, in other words, that the replication crisis started as an epistemological crisis but heralds a model crisis and possibly a paradigm shift. Such a paradigm shift in response to a fundamental problem also involves a different, new way of thinking, the emergence of an entirely different form of theorizing, and the need to develop new concepts that reflect this changed way of thinking—in short, a different *Weltanschauung* concerning the nature of the psyche.

## Author contributions

RM: Writing – original draft, Writing – review & editing, Conceptualization. IB: Writing – review & editing. JH: Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Agassi, J. (2014). Popper and his popular critics: Thomas Kuhn, Paul Feyerabend and Imre Lakatos. Cham: Springer.

Amrhein, V., Trafimow, D., and Greenland, S. (2019). Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *Am. Stat.* 73, 262–270. doi: 10.1080/00031305.2018.1543137

Anderson, S. F. (2020). Misinterpreting p: the discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychol. Methods* 25, 596–609. doi: 10.1037/met0000248

Anderson, S. F., and Maxwell, S. E. (2017). Addressing the "replication crisis": using original studies to design replication studies with appropriate statistical power. *Multivar. Behav. Res.* 52, 305–324. doi: 10.1080/00273171.2017.1289361

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *Eur. J. Personal.* 27, 108–119. doi: 10.1002/per.1919

Bachelard, G. (1974). Épistémologie. Textes choisis par dominique lecourt. Paris: Presses Universités de France.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454. doi: 10.1038/533452a

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., et al. (2018). Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10. doi: 10.1038/s41562-017-0189-z

Bird, A. (2013). "The historical turn in the philosophy of science" in The Routledge companion to philosophy of science. eds. M. Curd and S. Pillos (London: Routledge), 79–89.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl. Acad. Sci.* 119:e2203150119. doi: 10.1073/pnas.2203150119

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* 351, 1433–1436. doi: 10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644. doi: 10.1038/s41562-018-0399-z

Carter, J. A., and Littlejohn, C. (2021). This is epistemology: an introduction. New York, NY: John Wiley and Sons.

Chambers, C. (2017). The seven deadly sins of psychology: a manifesto for reforming the culture of scientific practice. Princeton, NJ: Princeton University Press.

Colling, L. J., and Szűcs, D. (2021). Statistical inference and the replication crisis. *Rev. Philos. Psychol.* 12, 121–147. doi: 10.1007/s13164-018-0421-4

Corr, P. J., DeYoung, C. G., and McNaughton, N. (2013). Motivation and personality: a neuropsychological perspective. *Soc. Personal. Psychol. Compass* 7, 158–175. doi: 10.1111/spc3.12016

Crump, T. (1992). The anthropology of numbers. Cambridge: Cambridge University Press.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect. Psychol. Sci.* 3, 286–300. doi: 10.1111/j.1745-6924.2008.00079.x

Derksen, M. (2019). Putting popper to work. *Theory Psychol.* 29, 449–465. doi: 10.1177/0959354319838343

Dolan, M., and Park, I. (2002). The neuropsychology of antisocial personality disorder. *Psychol. Med.* 32, 417–427. doi: 10.1017/S0033291702005378

Effingham, N. (2013). An introduction to ontology. New York, NY: John Wiley and Sons.

Elster, J. (2016). Tool-box or toy-box? Hard obscurantism in economic modeling. *Synthese* 193, 2159–2184. doi: 10.1007/s11229-015-0836-8

Eronen, M. I., and Bringmann, L. F. (2021). The theory crisis in psychology: how to move forward. *Perspect. Psychol. Sci.* 16, 779–788. doi: 10.1177/1745691620970586

Farrell, R. P. (2003). "Tightrope-walking rationality: Feyerabend's metanarrative" in Feyerabend and scientific values. Boston studies in the philosophy of science, Vol. 235. ed. I. J. Kidd (Dordrecht: Springer).

Feyerabend, P. (1975). Against method. Outline of an anarchistic theory of knowledge. London: New Left Books.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspect. Psychol. Sci.* 12, 46–61. doi: 10.1177/1745691616654458

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspect. Psychol. Sci.* 13, 433–438. doi: 10.1177/1745691617745651

Fisher, R. A. (1935/1974). The design of experiments. New York, NY: Macmillan Press.

Flanagan, O. (2007). The really hard problem. Meaning in a material world. London: MIT Press.

Flora, D. B. (2020). Thinking about effect sizes: from the replication crisis to a cumulative psychological science. *Can. Psychol.* 61, 318–330. doi: 10.1037/cap0000218

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspect. Psychol. Sci.* 7, 585–594. doi: 10.1177/1745691612459520

Fuller, S. (2003). Kuhn vs. Popper: the struggle for the soul of science. New York, NY: Columbia University Press.

Galak, J., LeBoeuf, R. A., Nelson, L. D., and Simmons, J. P. (2012). Correcting the past: failures to replicate psi. *J. Pers. Soc. Psychol.* 103, 933–948. doi: 10.1037/a0029709

Gelman, A., and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* 60, 328–331. doi: 10.1198/000313006X152649

Greenfield, P. M. (2017). Cultural change over time: why replicability should not be the gold standard in psychological science. *Perspect. Psychol. Sci.* 12, 762–771. doi: 10.1177/1745691617707314

Haig, B. D. (2022). Understanding replication in a way that is true to science. *Rev. Gen. Psychol.* 26, 224–240. doi: 10.1177/10892680211046514

Harris, C. R., Coburn, N., Rohrer, D., and Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS One* 8:e72467. doi: 10.1371/journal.pone.0072467

Henrich, J., Heine, S., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature* 466:29. doi: 10.1038/466029a

Horsten, L. (2023). "Philosophy of mathematics" in The Stanford encyclopedia of philosophy. ed. E. N. Zalta (Redwood City, CA: Stanford University Press).

Hutmacher, F., and Mayrhofer, R. (2023). Psychology as a historical science? Theoretical assumptions, methodological considerations, and potential pitfalls. *Curr. Psychol.,* 42, 18507–18514. doi: 10.1007/s12144-022-03030-0

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294, 218–228. doi: 10.1001/jama.294.2.218

Irvine, E. (2021). The role of replication studies in theory building. *Perspect. Psychol. Sci.* 16, 844–853. doi: 10.1177/1745691620970558

Jacquette, D. (2014). Ontology. London: Routledge.

Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., and Walsh, D. (2024). Naturalizing relevance realization: why agency and cognition are fundamentally not computational. *Front. Psychol.* 15:1362658. doi: 10.3389/fpsyg.2024.1362658

Juarrero, A. (2000). Dynamics in action: intentional behavior as a complex system. *Emergence* 2, 24–57. doi: 10.1207/S15327000EM0202_03

Keuth, H. (2005). The philosophy of Karl Popper. Cambridge: Cambridge University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability. *Soc. Psychol.* 45, 142–152. doi: 10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B. Jr., Alper, S., et al. (2018). Many labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* 1, 443–490. doi: 10.1177/2515245918810225

Kline, R. B. (2023). Principles and practice of structural equation modeling. New York, NY: Guilford Publications.

Koole, S. L., and Lakens, D. (2012). Rewarding replications: a sure and simple way to improve psychological science. *Perspect. Psychol. Sci.* 7, 608–614. doi: 10.1177/1745691612462586

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., et al. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Commun. Psychol.* 1:3. doi: 10.1038/s44271-023-00003-2

Kuhn, Thomas S. (1966). The structure of scientific revolutions. (3rd ed.). University of Chicago Press.

Lakatos, I. (1978). The methodology of scientific research programmes. Cambridge: Cambridge University Press.

Lakatos, I., and Musgrave, A. (1970). Criticism and the growth of knowledge. Cambridge: Cambridge University Press.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* 1, 259–269. doi: 10.1177/2515245918770963

Larvor, B. (1998). Lakatos: an introduction. London: Routledge.

Laws, K. R. (2016). Psychology, replication and beyond. *BMC Psychol.* 4, 30–38. doi: 10.1186/s40359-016-0135-2

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: righting the ship. *Perspect. Psychol. Sci.* 12, 660–664. doi: 10.1177/1745691616687745

Lilienfeld, S. O., and Strother, A. N. (2020). Psychological measurement and the replication crisis: four sacred cows. *Can. Psychol.* 61, 281–288. doi: 10.1037/cap0000236

Lycan, W. G. (2019). Philosophy of language. A contemporary introduction. London: Routledge.

Malich, L., and Rehmann-Sutter, C. (2022). Metascience is not enough – a plea for psychological humanities in the wake of the replication crisis. *Rev. Gen. Psychol.* 26, 261–273. doi: 10.1177/10892680221083876

Marcum, J. A. (2005). Thomas Kuhn's revolution: An historical philosophy of science. London: Continuum.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70, 487–498. doi: 10.1037/a0039400

Mayrhofer, R., and Hutmacher, F. (2020). The principle of inversion: why the quantitative-empirical paradigm cannot serve as a unifying basis for psychology as an academic discipline. *Front. Psychol.* 11:596425. doi: 10.3389/fpsyg.2020.596425

Morawski, J. (2019). The replication crisis: how might philosophy and theory of psychology be of use? *J. Theor. Philos. Psychol.* 39, 218–238. doi: 10.1037/teo0000129

Muñoz, J., and Young, C. (2018). We ran 9 billion regressions: eliminating false positives through computational model robustness. *Sociol. Methodol.* 48, 1–33. doi: 10.1177/0081175018777988

Musgrave, A. (1993). Popper on induction. *Philos. Soc. Sci.* 23, 516–527. doi: 10.1177/004839319302300407

Nickles, T. (2003). Thomas Kuhn on revolution and Paul Feyerabend on anarchy. Cambridge: Cambridge University Press.

Nagel, T. (1974). What Is It Like to Be a Bat?. *Philos. Rev.,* 83, 435–450.

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* 26, 1596–1618. doi: 10.3758/s13423-019-01645-2

Oberheim, E. (2006). Feyerabend's philosophy. Berlin: De Gruyter.

O'Donohue, W. T., Masuda, A., and Lilienfeld, S. O. (2022). Avoiding questionable research practices in applied psychology. Cham: Springer.

Open Science Collaboration (2012). An open, large scale, collaborative effort to estimate the reproducibility of psychological science. Perspectives on psychological. *Science* 7, 657–660. doi: 10.1177/1745691612462588

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Front. Psychol.* 10:2970. doi: 10.3389/fpsyg.2019.02970

Perović, K. (2024). Ontological categories. A methodological guide. Cambridge: Cambridge University Press.

Platts, M. (1997). Ways of meaning. An introduction to philosophy of language. *2nd* Edn. London: MIT Press.

Popper, K. (1959/2005). The logic of scientific discovery. London: Routledge.

Preston, J. (1997). Feyerabend: philosophy, science and society. Cambridge: Polity Press.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philos. Compass* 14:e12633. doi: 10.1111/phc3.12633

Rowbottom, D. P. (2011). Popper's critical rationalism. A philosophical investigation. London: Routledge.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a metanalytic review of the testing effect. *Psychol. Bull.* 140, 1432–1463. doi: 10.1037/a0037559

Rowlands, M. (2001). The nature of consciousness. Cambridge: Cambridge University Press.

Sandkühler, H. J. (1991). Die wirklichkeit des wissens: geschichtliche einführung in die epistemologie und theorie der erkenntnis. Berlin: Suhrkamp.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant Child Dev.* 31:e2295. doi: 10.1002/icd.2295

Scheel, A. M., Tiokhin, L., Isager, P. M., and Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspect. Psychol. Sci.* 16, 744–755. doi: 10.1177/1745691620966795

Schmidt, F. L., and Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: is lack of replication the real problem? Or is it something else? *Arch. Sci. Psychol.* 4, 32–37. doi: 10.1037/arc0000029

Schretlen, D. J., van der Hulst, E. J., Pearlson, G. D., and Gordon, B. (2010). A neuropsychological study of personality: trait openness in relation to intelligence, fluency, and executive functioning. *J. Clin. Exp. Neuropsychol.* 32, 1068–1073. doi: 10.1080/13803391003689770

Schwieren, J., Barenberg, J., and Dutke, S. (2017). The testing effect in the psychology classroom: a meta-analytic perspective. *Psychol. Learn. Teach.* 16, 179–196. doi: 10.1177/1475725717695149

Serra-Garcia, M., and Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Sci. Adv.* 7:eabd1705. doi: 10.1126/sciadv.abd1705

Shapiro, S. (2000). Thinking about mathematics: the philosophy of mathematics. Oxford: Oxford University Press.

Sharpe, D., and Poets, S. (2020). Meta-analysis as a response to the replication crisis. *Can. Psychol.* 61, 377–387. doi: 10.1037/cap0000215

Shenal, B. V., Harrison, D. W., and Demaree, H. A. (2003). The neuropsychology of depression: a literature review and preliminary model. *Neuropsychol. Rev.* 13, 33–42. doi: 10.1023/A:1022300622902

Shrout, P. E., and Rodgers, J. L. (2018). Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* 69, 487–510. doi: 10.1146/annurev-psych-122216-011845

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., et al. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* 1, 337–356. doi: 10.1177/2515245917747646

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nat. Hum. Behav.* 4, 1208–1214. doi: 10.1038/s41562-020-0912-z

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* 26, 559–569. doi: 10.1177/0956797614567341

Soteriou, M. (2018). "The ontology of emotion" in The ontology of emotions. eds. H. Naar and F. Teroni (Cambridge: Cambridge University Press), 71–89.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* 11, 702–712. doi: 10.1177/1745691616658637

Steup, M. (1996). An introduction to contemporary epistemology. New York, NY: Simon and Schuster.

Steup, M., and Ram, N. (2024). "Epistemology" in The Stanford encyclopedia of philosophy. eds. E. N. Zalta and U. Nodelman (Redwood City, CA: Stanford University Press).

Stroebe, W., and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9, 59–71. doi: 10.1177/1745691613514450

Swann, A. J. (1988). Popper on induction. *Br. J. Philos. Sci.* 39, 367–373. doi: 10.1093/bjps/39.3.367

Toulmin, S. (1972). Human understanding. Oxford: Clarendon Press.

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psychol.* 37, 1–2. doi: 10.1080/01973533.2015.1012991

Tye, M. (2021). "Qualia" in The Stanford encyclopedia of philosophy. eds. E. N. Zalta and U. Nodelman (Redwood City, CA: Stanford University Press).

Uher, J. (2021). Psychology's status as a science: peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integr. Psychol. Behav. Sci.* 55, 212–224. doi: 10.1007/s12124-020-09545-0

Verhagen, J., and Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* 143:1457. doi: 10.1037/a0036731

Weir, R. S. (2024). The mind-body problem and metaphysics: an argument from consciousness to mental substance. London: Routledge.

Westerhoff, J. (2005). Ontological categories: their nature and significance. Oxford: Clarendon Press.

Wiggins, B. J., and Christopherson, C. D. (2019). The replication crisis in psychology: an overview for theoretical and philosophical psychology. *J. Theor. Philos. Psychol.* 39, 202–217. doi: 10.1037/teo0000137

Young, C. (2018). Model uncertainty and the crisis in science. *Socius* 4:237802311773720. doi: 10.1177/2378023117737206

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 02 October 2024
DOI 10.3389/fpsyg.2024.1399317



OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
Robert Mislevy,
University of Maryland, United States
Andreas Sudmann,
University of Bonn, Germany

*CORRESPONDENCE
Alex Scharaschkin
✉ ascharaschkin@aqa.org.uk

RECEIVED 11 March 2024
ACCEPTED 19 September 2024
PUBLISHED 02 October 2024

CITATION
Scharaschkin A (2024) Educational assessment without numbers. *Front. Psychol.* 15:1399317. doi: 10.3389/fpsyg.2024.1399317



# Educational assessment without numbers

Alex Scharaschkin[1,2]*

[1]Department of Education, University of Oxford, Oxford, United Kingdom, [2]AQA Education, London, United Kingdom

Psychometrics conceptualizes a person's *proficiency* (or *ability*, or *competence*), in a cognitive or educational domain, as a latent numerical quantity. Yet both conceptual and empirical studies have shown that the assumption of quantitative structure for such phenomena is unlikely to be tenable. A reason why most applications of psychometrics nevertheless continue to treat them as if they were numerical quantities may be that quantification is thought to be necessary to enable *measurement*. This is indeed true if one regards the task of measurement as the location of a measurand at a point on the real number line (the viewpoint adopted by, for example, the representational theory of measurement, the realist theory of measurement as the discovery of ratios, and Rasch measurement theory). But this is not the only philosophically respectable way of defining the notion of measurement. This paper suggests that van Fraassen's more expansive view of measurement as, in general, *location in a logical space* (which could be the real continuum, as in metrological applications in the physical sciences, but could be a different mathematical structure), provides a more appropriate conceptual framework for psychometrics. Taking educational measurement as a case study, it explores what that could look like in practice, drawing on fuzzy logic and mathematical order theory. It suggests that applying this approach to the assessment of intersubjectively constructed phenomena, such as a learner's proficiency in an inherently fuzzily-defined subject area, entails recognizing the theory-dependent nature of valid representations of such phenomena, which need not be conceived of structurally as values of quantities. Finally, some connections are made between this "qualitative mathematical" theorization of educational assessment, and the application of techniques from machine learning and artificial intelligence in this area.

KEYWORDS

**theory and philosophy of measurement, psychometrics, educational assessment, van Fraassen, qualitative mathematics, concept lattice, fuzzy logic**

## 1 Introduction

The question of what it could mean to *measure* phenomena that form the basis of theory and debate in the human sciences, such as human attitudes, opinions, dispositions, or psychological or cognitive traits, has been a subject of critical enquiry since at least the mid eighteenth century (Michell, 1999). For example, the question of whether such phenomena could be *quantified* was contested by Reid (1849), even before a clearer definition of "a quantity" had been put forward by Hölder (1901).

This paper considers the question of measuring educational constructs, such as a learner's *ability*, or *proficiency*, or *competence* in a subject, field of study, or educational domain. Many educational tests and assessment procedures—some of them used to make high-stakes decisions about the test-takers—apparently produce, or claim to produce, numerical measurements of such properties, such that learners can be placed on a

quantitative *scale* with respect to them. Psychometrics is the application of statistical methods to the study of psychological and educational phenomena. It relies on the particular mathematical characteristics of quantitative structures (in practice, the real numbers and vector spaces over the reals) to perform calculations and procedures that are used as the warrants for substantive conclusions, such as "how much" ability a student is estimated to have, or how to equate measurements of ability derived from different tests.

The paper argues that the reliance of psychometrics on quantitative structures is grounded in an assumption that *quantification* is necessary to allow *measurement*. It proposes, however, that psychological and educational measurement need not be reliant on numbers. It suggests that van Fraassen's (2008) account of measurement as a process whereby the measurand is located in an appropriate "logical space" is well-suited to serve as a foundation for an account of the measurement of educational phenomena such as students' abilities or competencies in a subject domain—phenomena that are arguably inherently "fuzzy" and multifaceted. Such a logical space *could* be the particular mathematical structure that uniquely characterizes the real numbers (a complete ordered field, in mathematical terminology), but it need not be.

The structure of the paper is as follows. Section 2 briefly outlines the approach to measuring cognitive and educational constructs, by assuming quantitative structure, that became standard in psychometrics over the twentieth century. It summarizes critiques of the quantity assumption, and argues that these critiques have sufficient conceptual and empirical weight to warrant a serious explanation of what an approach to psychological and educational measurement could look like if the assumption is set aside. Taking the example of summative educational assessment in particular, it suggests that in many cases construct validity may be better served by a more generalized view of measurement, of the kind proposed by van Fraassen (2008). Van Fraassen's approach is explained in more detail in Section 3.

Section 4 makes the discussion more concrete by comparing quantitative and qualitative measurement approaches for a toy example of an educational test. This is extended in Section 5 to a consideration of the practicalities—in particular, the computational complexity—of applying qualitative mathematical (fuzzy order-theoretic) methods to the kinds of test response data that arise in real practice. And since traditional methods of analysis of educational assessment data are increasingly being supplemented, or even supplanted, by the application of techniques from natural language processing, machine learning, and artificial intelligence (AI), Section 6 considers some of the connections between educational measurement and AI-enabled classification procedures. Finally, the concluding discussion in Section 7 poses some questions for further research. It concludes that it is worth pursuing further conceptual and technical development of non-quantitative measurement approaches in psychometrics, especially since, with the rapid rise and application of AI (e.g., von Davier et al., 2021), there is a risk that psychometrics is simply replaced with data science—with the loss of substantive theoretical content concerning construct definition and the design of valid measurement procedures. A way forward is for psychometrics itself to develop into a discipline that rests on quantitative measurement when it is appropriate, but does not exclude a broader view.

# 2 Quantification in psychometrics

## 2.1 Abilities as latent quantities

Psychometrics normally conceptualizes a learner's *ability* (or *proficiency*, or *competence*) in a domain as a latent numerical quantity, $\theta$ (Kline, 2000; van der Linden and Hambleton, 1997). For each learner, a value of $\theta$ is calculated from the observed data arising from an assessment (e.g., item response data). The "more $\theta$" a learner has (the higher their value of $\theta$), the "better at" the assessment construct they are taken to be (modulo some "measurement error"). That is to say, the relation of *betterness*, between learners, as to the different levels, states, or configurations of their abilities, is taken to be adequately captured by the relation of *order* ($\geq$) between numerical values. Moreover, to allow a value of $\theta$ actually to be derived for each learner, the set of all possible $\theta$-values is normally supposed not only to be totally ordered, but quantitative and continuous.[1] Making these structural assumptions about the property of *ability* enables it to be treated as if it were a real number. Hence the whole array of statistical techniques whose mathematical validity depends on the metric and topological properties of the real numbers (such as factor analysis, item response theory, maximum likelihood estimation, etc.) can be applied to obtain numerical values that are taken to be *measurements of learners' abilities* in the cognitive or educational domain in question.

This paper will argue that one should not think of the "betterness" relation between learners, as to their proficiency in a particular educational domain, as a total order relation (a ranking), in general, but rather as a partial order.[2] Sometimes the way in which the assessment construct is defined will allow learners to be ranked as to their proficiency with respect to that construct. In other cases, it may only be possible to infer, for some pairs of learners, that their proficiency states, or levels, are non-comparable (qualitatively different). This does not preclude the possibility of

1 See the Appendix for definitions of *total order* and *quantity*. Informally, a totally ordered set $X$ is one in which all the members can be ranked—there is an ordering $\geq$ such that either $x \geq y$ or $y \geq x$, for all $x$ and $y$ in $X$. A property is a quantity if its values are totally ordered and also additive—that is, they can be combined in a way that mirrors the properties of the addition of numbers. Additivity is required for a property's values to form an *interval scale* or a *ratio scale*, in the terminology of Stevens (1946). A quantitative property is *continuous* if its possible values form a continuum with no "gaps".

2 See the Appendix for a formal definition of *partial order*. In essence, when entities are partially ordered, there may exist pairs of entities that are not directly comparable, and the entities cannot necessarily be placed in a single linear sequence (a ranking) with respect to the feature of interest. In educational tests, each individual item (question or task) typically totally orders the respondents with respect to that item (for example "those who got the question right" $\geq$ "those who got the question wrong"; or "those who scored 3 marks" $\geq$ "those who scored 2 marks" $\geq$ "those who scored 1 mark' $\geq$ "those who scored 0 marks") In general, however, the joint result (the product) of all of these total orders is an overall partial ordering of respondents, with some patterns of item responses not being directly comparable with others.

grouping learners together into "coarser" ordinal classes (such as examination grades), such that one can infer that those who "pass" are more proficient than those who "fail", for instance. It just means that, within the "pass" category, there may be some learners whose proficiencies, although both of at least a "pass" level, may be different, and non-comparable. This argument is developed further in Section 4 below.

There is a literature that critically examines the plausibility of assuming quantitative structure for phenomena such as ability (for example, Michell, 2006, 2009, 2012, 2013; Heene, 2013; Kyngdon, 2011; McGrane and Maul, 2020, and from a broader perspective, Uher, 2021, 2022a). One focus of this has been what Michell (2012) calls the "psychometricians' fallacy": the implicit leap that is often made, from maintaining that a property has a totally-ordered structure (that its possible values, states, or levels can be ranked, that is, placed on an *ordinal scale*, as described by Stevens, 1946), to treating it as if it had quantitative structure (as if its values formed an *interval* or a *ratio* scale, in Stevens' typology).

In some cases it is possible to test empirically whether a property whose values are ordered is plausibly likely to have the further structure required for it to be quantitative. This is discussed in Section 2.2.2. Yet at an even more basic level, one might question why a construct such as *ability* with respect to a given cognitive or educational domain (specified in a more-or-less precise way), should even be regarded as a property that necessarily ought to have a totally ordered structure. Must it be a phenomenon that only occurs in such a way that any one person's ability-state is always linearly comparable with (larger than, the same as, or smaller than) any other person's state? Uher (2022b) makes an analogous point with respect to the use of rating scales to "measure" the property of agreement.

If one considers the actual data upon which the inferences derived from educational testing procedures are based, then as Kane (2008) notes, "we are likely to have, at best, a partial ordering, unless we arbitrarily decide that some patterns [of item response] are better than others". In practice, and as discussed further in Section 4, almost all psychometric approaches to working with such partially-ordered data do indeed involve making decisions about how to use the data to generate a total order (with each learner's score being their location with respect to this total order).

The question whether such decisions are indeed "arbitrary" (and if not, which one is best or most appropriate) hinges, again, on how the measurand—each respondent's ability in the domain in question—is conceptualized. This issue is well-described by Maul (2017, p. 60), who notes that

> Any effort to construct a measure of an attribute will have trouble getting off the ground in the absence of a sufficiently well-formed definition of the target attribute, including an account of what it means for the attribute to vary (i.e., what meaning can be attached to claims about there being "more" or "less" of it, between and possibly within individuals) and how such variation is related to variation in the observed outcomes of the instrument (i.e., item response behaviour).

It is suggested in Section 3.2 that questions of this kind form part of what van Fraassen (2008) refers to as the *data model* for the target attribute. It is rather rare for psychometrics textbooks to devote much attention to these theoretical or conceptual issues, however. Often (e.g., Raykov and Marcoulides, 2011) it is stated that psychological and educational measurement is concerned with appraising how individuals differ with regard to hypothesized, but not directly observable, attributes or traits, such as intelligence, anxiety, or extraversion. It is assumed that these traits are in fact quantities (for instance Kline, 2000, p. 18) simply states that "the vast majority of psychological tests measuring intelligence, ability, personality and motivation … are interval scales"), and models are then introduced to relate them to observable data such as test or questionnaire responses in such a way as to enable the numerical latent trait parameters to be estimated, together with measures of precision such as standard errors—all conditional on the adequacy and plausibility of the model that has been assumed. Of course if the model is not adequate as a structural theory of the phenomenon itself, then results may simply reflect artifacts of the model (e.g., consequences—sometimes rather trivial tautologies—that follow from the metric structure of the real numbers), rather than corresponding to valid inferences with respect to the theory of the phenomenon.

Why should a phenomenon such as a learner's proficiency or competence in a particular domain be assumed to have the structure of a total order (let alone a quantity)? The reason probably goes back to a belief fundamental to the early development of psychometrics, that quantitative structure is necessary to enable measurement. For example, Thurstone (1928) claimed that

> When the idea of measurement is applied to scholastic achievement, … it is necessary to force the qualitative variations [in learners' performances] into a quantitative linear scale of some sort.

If "the idea of measurement" entails *locating a measurand at a point on the real number line*, then "forcing" observed qualitative variations to fit a quantitative structure is an understandable approach to adopt (even if it raises questions about validity). Indeed two common theoretical frameworks for psychological and educational measurement—the representational theory of measurement, and Rasch measurement theory—could be construed as concerned with ways to "force" qualitative variation into quantitative form: the former by aiming to define conditions under which qualitative observations can be mapped into numerical structures; the latter by rejecting observations that do not fit an assumed quantitative model. These approaches are unpacked a little in the next section.

## 2.2 Theories of measurement

### 2.2.1 The representational theory of measurement

Tal (2020), in his survey of the philosophy of measurement in science, describes the representational theory of measurement (RTM) as "the most influential mathematical theory of measurement to date". Wolff (2020), in a recent structuralist account of quantity and measurement, calls it "arguably the most developed formal theory of measurement". Michell (1990) claimed that it is "the orthodox theory of measurement within the philosophy of science".

The canonical text on RTM (Krantz et al., 1971, p. 9) takes *measurement* to mean "the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful".

RTM supposes that we are given an "empirical relational structure" (itself an abstraction of certain features of an "observed reality"). This structure consists of objects, relations between them, and possibly also ways of combining or composing them. For example in educational measurement contexts, we might take as objects students' responses to a writing task, and consider a binary relation $\succeq$ of *betterness* as being of interest (as in "student *X*'s piece of writing is a better response to the task than student *Y*'s: $X \succeq Y$"). Or we might be interested in how parts of a test or assessment combine (via a binary operation $\bullet$) to form an overall measure. For example, "correctly answering questions 3 and 4 demonstrates a higher level of proficiency than correctly answering questions 1 and 2": $q_3 \bullet q_4 \succeq q_1 \bullet q_2$. We might then wish to investigate whether these aspects of students' responses to tasks—this empirical relational structure—can be mapped to a numerical ordering or scoring system, in such a way that the structure is preserved (e.g., relative betterness between responses is mirrored by the relative magnitudes of the numbers assigned to those responses).

The idea is that if such homomorphisms can be shown to exist, then inferences in the numerical relational structure (normally taken to be the real numbers with the usual order relation $\geq$ and binary operations $+$ and $\cdot$) provide warrants for conclusions in the substantive domain of the empirical relational structure. If, further, we posit that differences in the observed outcomes of an educational assessment procedure, such as the administration of a test or examination, are *caused by* differences in the configurations, between learners, of their "underlying proficiency", then establishing a homomorphism between the empirical relational structure and the real numbers [i.e., establishing that the outcomes can be "placed on an interval (or ratio) scale"] serves to justify the assumption of quantitative structure for this assumed underlying proficiency trait, and hence to enable the measurement of each test-taker's proficiency by locating them at the point on the real line that corresponds to their level of proficiency.

### 2.2.2 Qualitative relational structures and testing for quantity

The adequacy of RTM as a theory of measurement has been extensively critiqued (see, e.g., Michell, 1990, 2021; see also Luce and Narens, 1994), with commentaries noting that its abstract nature sidesteps the actual process of measuring anything, the construction of measuring instruments, and any discussion of measurement error. The merits of such critiques are not discussed further in this paper, because the position adopted here will be that of Heilmann (2015). Heilmann (2015, p. 789) does not assess RTM as a candidate for a theory of measurement, but rather as a collection of mathematical theorems: theorems whose structure makes them useful for investigating problems of concept formation. He proposes viewing theorems in RTM as

> providing us with mathematical structures which, if sustained by specific conceptual interpretations, can provide insights into the possibilities and limits of representing concepts numerically

He regards RTM as studying not mappings from an empirical relational structure to a numerical relational structure, but rather from a *qualitative relational structure* (QRS) to a numerical relational structure. Taken in that sense, he argues, RTM can provide tools for testing the extent to which abstract concepts (captured or described as qualitative relational structures) can be represented numerically.[3]

Arguably, this is how RTM (including in particular the subset of RTM theorems that form the so-called theory of *conjoint measurement*: see Luce and Tukey, 1964) does in fact tend to be used in the literature exploring the plausibility of assuming quantitative structure for educational, psychological, or social measurands.

For example, Michell (1990) re-analyzed data collected by Thurstone (1927b) regarding judgements as to the seriousness of various crimes. Thurstone (1927a) claimed that his theory of *comparative judgement* enabled the construction of a *quantitative scale* for the measurement of seriousness of crime, by applying the theory to the outcomes of a collection of pairwise comparisons, in which subjects were repeatedly asked which of two crimes presented to them was the more serious. Michell (1990, p. 107) carefully stated the assumptions of Thurstone's theory, and demonstrated by applying results from RTM that "either seriousness of crimes is not a quantitative variable, or else some other part of Thurstone's theory of comparative judgement is false".

van Rooij (2011) applied theorems from RTM to explore whether properties of objects, that manifest linguistically as adjectives with comparative degrees, can be represented numerically, what scale properties may hold for them, and hence whether inter-adjective comparisons (such as "*x* is *P*-er than *y* is *Q*") can be meaningful. This is analogous to the vexed question, in educational assessment, of inter-subject comparison when it comes to setting and maintaining qualification standards (see, e.g., Newton et al., 2007; Coe, 2008).

Karabatsos (2001, 2018), Kyngdon (2011), Domingue (2014), and Scharaschkin (2023) applied theorems from RTM to the question of testing whether psychometric attributes comply with requirements for quantitative structure, combining the RTM results with a stochastic approach to address expected "measurement error" in most measurement scenarios with reasonable numbers of test-takers and test items. Domingue found that the results of a well-known test of reading showed that it was highly implausible that reading proficiency was a quantitatively-structured variable. Scharaschkin found that the results of a test of physics for school-leavers did not support the assumption of quantitative structure

3 A further extension of Heilmann's position would be to consider mappings from a QRS to another QRS: in other words, to relax the restriction that the "representing" structure should be numerical. Such a generalization might permit both RTM and van Fraassen's approach to be located, from a formal mathematical perspective, within the general theory of structure known as category theory, but will not be pursued here.

for a hypothesized "physics proficiency" construct. On the other hand, he found that the results of a similar test of economics were approximately consistent with an assumption of quantitative structure.

None of these applications require assuming the validity or adequacy of RTM as a substantive theory of measurement—indeed, Michell (2021) explicitly rejects it. Yet they do shed light on the extent to which qualitatively-structured data can be treated *as if* it were a manifestation of quantitatively-structured latent traits, and provide empirical evidence that it is not always valid to do so.

This is relevant to the practice of educational assessment and test construction because most practitioners and test developers probably do work within a pragmatic "as if" framework, as summarized by Lord and Novick (1968, p. 358):

> Much of psychological theory is based on trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behaviour.

Some of the ways in which theories of cognition have been more directly incorporated into the use of quantitative latent variable modeling, and their relation to the ideas considered in this paper, are discussed further in Section 5.4.

### 2.2.3 Rasch measurement theory

Psychometrics conducted in the Rasch measurement tradition (Andrich and Marais, 2019) takes the view that measurement is only meaningful for quantitative phenomena. Thus, if a putative measurement procedure such as an educational or psychological test yields results that are inconsistent with a underlying quantitative variable, then the procedure is not, in fact, *bona fide* measurement, and requires modification. In practice this means modifying tests by deleting or changing items until a sufficiently good fit to the Rasch model is obtained.[4]

So rather than trying to find a model that fits the data that has been obtained from the administration of a test, the Rasch measurement approach is to try to make the data fit the model. Modifying the measurement instrument to achieve this may come at the cost of severely constraining the theory of (or, in the terminology of Section 3.2, the relevant data model for) the substantive phenomenon or construct of interest. It might be that the construct cannot be sufficiently constrained or re-defined without significantly departing from its underpinning theory of value. In an educational assessment context, this would be the case if making such changes to the assessment instrument would compromise construct validity: the assessors' understanding of what constitute the key attributes of proficiency in the given domain, and how relatively better/worse/different states of proficiency would present with respect to these attributes. In such cases the choice would seem to be either to abandon the idea of measuring the construct at all, or to abandon the restriction of measurement to locating measurands within solely quantitative mathematical structures. This paper explores the latter option.

### 2.2.4 Measurements as ratios

Michell (1999) traces the evolution of the concept of measurement in psychology since the publication of Fechner's *Elemente der Psychophysik* in 1860. He bemoans the movement away from the conceptualization of measurement that had become standard in nineteenth century physics, namely (Michell, 1999, p. 14) "the discovery[5] or estimation of the ratio of the magnitude of a quantitative attribute to a unit (a unit being, in principle, any magnitude of the same quantitative attribute)". In other words, as elementary physics texts still state, physical quantity = real number × unit, where the real number is the measurement of the physical quantity.

Michell notes (p. 19) that "according to the traditional understanding of measurement, only attributes which possess quantitative structure are measurable. This is because only quantitative structure sustains ratios". He argues that, this being the case, it is incumbent on psychometricians to investigate whether the phenomena they study do, in fact, have quantitative structure, before applying statistical models that assume it. Since in practice this is almost never done, his claim is that, for the most part, "psychometrics is built upon a myth" (Michell, 2012). Once again, the choice appears to be to accept the constraints of the "traditional understanding of measurement", or to explore whether psychometrics could benefit from engagement with a more expansive conceptualization of what it means to measure something. The next section considers such a viewpoint.

## 3 van Fraassen's account of measurement

### 3.1 Basic principles and relevance to psychometrics

Bas van Fraassen's (2008) *Scientific Representation: Paradoxes of Perspective* is an empiricist structuralist account of measurement and representation in science. This stance eschews debate about the ontological status of the phenomena or reality that scientific theories describe, and concerns itself rather with elucidation of

4 The Rasch model, also known as the 1-parameter item response model, postulates that the log-odds of a test-taker of ability $\theta$ correctly answering an item of difficulty $\delta$ is simply $\theta - \delta$ (in the case of a test consisting of a sequence of dichotomously-scored items). There are of course other item response models that postulate additional item parameters, but Rasch theorists hold that the 1-parameter model is theoretically more appropriate as a basis for enabling measurement because it enables, within a given collection of persons and items, so-called invariant comparisons of persons (as to their ability) and items (as to their difficulty): see Andrich and Marais (2019, p. 80).

5 The development of quantum theory in the twentieth century problematized the classical epistemological viewpoint on measurement as "discovery". As Peres (1995, p. 14) observes, "classical physics assumes that the property which is measured objectively *exists* prior to the interaction of the measuring apparatus with the observed system. Quantum physics, on the other hand, is incompatible with the proposition that measurements discover some unknown but pre-existing reality."

what van_fraassen argues is the key aim of developing and testing such theories, namely their empirical adequacy. van Fraassen (2008, p. 2) claims that "measuring, just as well as theorizing, is representing … measuring *locates* the target in a theoretically constructed logical space". To be more precise (p. 164),

> measurement is an operation that locates an item (already classified in the domain of a given theory) in a logical space (provided by the theory to represent a range of possible states or characteristics of such items).

A key point here is the theory-relatedness of measurement procedures. Echoing Maul's (2017) requirements, quoted in Section 2.1, for a "well-formed definition of the target attribute" as fundamental to psychometric measurement, van Fraassen suggests (p. 166) that "once a stable theory has been achieved, the distinction between what is and is not genuine measurement will be answered *relative to that theory*".

It is argued in Section 4 that a candidate theory for the phenomena (proficiency or competence in a domain) that form the subject matter of educational measurement, is a description of what constitutes betterness between learners' possible states or configurations of proficiency in a given domain. "Betterness"—which, as noted in Section 2, may be a more general order relation than a simple ranking—has to be defined in terms of criteria that may, in general, be manifested with *fuzzy degrees of truth* in the responses of learners to tasks that have been designed to provide information about their proficiency in the domain in question.

van Fraassen considers several measuring procedures in classical and quantum physics (p. 157–172 and 312–316), and concludes (p. 172) that they are all "cases of grading, in a generalized sense: they serve to classify items as in a certain respect greater, less, or equal. But … this does not establish that the scale must be the real number continuum, nor even that the order is linear. The range may be an algebra, a lattice, or even more rudimentary, a poset". In fact, Section 4 below considers the case of lattices as logical spaces for educational measurement procedures.[6]

It is worth exploring how van Fraassen's approach could be applied to educational measurement for at least two reasons. Firstly because, as discussed in Section 2.2.2, the mathematically necessary conditions for a learner's proficiency in a given educational domain to have the structure of a quantity often do not hold; and it is not possible to massage the assessment instrument to make them hold without loss of construct validity. In such cases, it would arguably be inappropriate to theorize the construct as quantitative, and hence its measurement as location *on the real line*, rather than in some other, theory-relevant, logical space.

Secondly, the approach of thinking about educational assessment constructs in terms of fuzzy criteria of value (what will count as creditworthy, or indicative of good/bad performance, in relation to what particular domain content) is what *actually happens in practice*, when subject domain experts develop and administer at least one kind of high-volume, high-stakes, educational assessment procedure, namely the public examinations taken by school pupils aged 16 and 18 in the UK. This brings us to a consideration of what van Fraassen calls *data models*.

## 3.2 Data and surface models

Measurements arise from the results of procedures designed to gather information about a phenomenon of interest. As noted in Section 2.2.2, these entail selective attention to specific features that are deemed to be relevant. That is to say, measuring a phenomenon involves collecting data structured in a specific way. van Fraassen (2008, p. 253) calls such a structure a *data model* for the measurand in question. He notes that

> A data model is relevant for a given phenomenon, not because of any abstract structural features of the model, but because it was constructed on the basis of results gathered in a certain way, selected by specific criteria of relevance, on certain occasions, in a practical experimental or observational setting designed for that purpose.

In educational measurement we have gathered in a certain way (via an assessment procedure such as a test), selected by specific criteria of relevance (construct-relevant criteria: Pollitt and Ahmed, 2008) on certain occasions (at a particular point or points in time), in a practical setting designed for that purpose (e.g., the rules of administration and physical requirements for conducting an examination).

In the case where the test consists of a sequence of dichotomously-scored items $I := \{i_1, \ldots, i_n\}$ administered to a collection $L := \{l_1, \ldots, l_m\}$ of learners, we can think of this measurement setup as a map $V : L \times I \to \{0, 1\}$ that assigns to each instance of a learner encountering an item the valuation 1 if they answer it correctly, and 0 if they answer it incorrectly. Equivalently, we can think of the information collected by the assessment procedure as organized in an $m \times n$ matrix whose $(m, n)$ entry is $V(l_m, i_n)$. There is, however, more structure entailed by the "betterness" ordering within each item (namely that "1" is better than "0") than immediately stands out from simply viewing the data as a table. As discussed in Section 4.2, the totality of the results-plus-valuation-system can be viewed as a lattice (the so-called *concept lattice* for the data table)—and it is suggested in Section 4 that such lattices (generalized to incorporate fuzzy valuations if necessary) form the natural data model for the phenomena that educational measurement procedures, such as tests and examinations, aim to measure.

van Fraassen (2008, p.253) describes constructing a data model as "precisely the selective relevant depiction of the phenomena *by the user of the theory* required for the possibility of representation of the phenomenon." In the context of educational testing, the proficiencies being studied are proficiencies or competencies *with respect to* a specified domain (such as "high school chemistry", or "A level French"). What "good performance" or "good demonstrated attainment" looks like in these domains (and hence what would count as evidence of better or worse levels, or states, or configurations, of learners' *proficiencies*) is always subject to a prevailing understanding or agreement as to what potential aspects

6 Algebras, lattices, and posets (short for partially-ordered sets) are types of mathematical structures. In particular, a lattice is a partially-ordered set (see the Appendix for a definition) in which each pair of elements has a least upper bound and a greatest lower bound.

of the domain are chosen as relevant for discrimination between learners' performances as to their quality. In other words, the criteria for creditworthiness of candidates' responses to tasks in an assessment can be regarded as the selective relevant depiction of the phenomenon of interest, by those members of the competent authority (the "users of the theory") who design, administer, and grade the tests. For that reason, concept lattices derived from the outcome data from the tests, that encode the relationship between learners and the assessment criteria, are appropriate data models.

In practice, van Fraassen (2008, p.167) notes that data models may be "abstracted into a mathematically idealized form" before empirical or experimental results are used to explore theories or explanations, or for substantive purposes. He gives the example of a data model consisting of relative frequencies, which is "smoothed" such that frequency counts are replaced with probabilities. An idealized or simplified version of a data model is called a *surface model* for the phenomenon in question. Surface models are considered further in Section 5.

# 4 Theories of constructs: comparing item response theory and fuzzy concept analysis

## 4.1 A small example

Table 1 shows results from an assessment that generates data on each of three items (or attributes) $\{i_1, i_2, i_3\}$ for six learners $\{l_1, \ldots, l_6\}$. Here 0 means "not demonstrated", $\frac{1}{2}$ means "partially demonstrated", and 1 (or $\frac{2}{2}$) means "fully demonstrated".

A traditional psychometric approach to analyzing this kind of data would be to treat each learner's results from the assessment as a vector in $\mathbb{R}^3$, and each learner's proficiency measure as a quantity (a point in $\mathbb{R}$). For example, we could treat the label for each item response category as a number, and add them to get a total score for each learner. This orders learners, with respect to proficiency, equivalently to fitting a Rasch model (a 1-parameter item-response model), since total score is a sufficient statistic for estimating proficiency in this model. Or we could do a principal components analysis and take the projection of each learner's item-response vector onto the component that accounts for the most variance as their proficiency measure (this is equivalent to fitting a 2-parameter item-response model: see Cho, 2023). Doing so for the data in Table 1 yields three components of which the first accounts for 72% of the variance in outcomes, with the other two accounting for 19 and 9%, respectively. We could therefore take the loading (projection) of each learner's results onto the first component as their score on an "underlying" quantitative variable that represents the assessment construct reasonably well. Figure 1 shows how learners' proficiency measures differ depending on the approach taken.

However, in view of the problems associated with assuming quantitative structure for proficiency discussed in Section 2.1 (tantamount, in Section 3.2's terms, to replacing the data model with a radically different surface model), let us consider a non-quantitative approach. If we take each learner's test response not as a vector of numbers, but rather a vector of ordered labels, then

TABLE 1 Data from a test.

| \ | $i_1$ | $i_2$ | $i_3$ |
|---|---|---|---|
| $l_1$ | 0 | $\frac{1}{2}$ | $\frac{1}{2}$ |
| $l_2$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ |
| $l_3$ | 1 | 1 | $\frac{1}{2}$ |
| $l_4$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ |
| $l_5$ | 0 | $\frac{1}{2}$ | 0 |
| $l_6$ | $\frac{1}{2}$ | 1 | 1 |

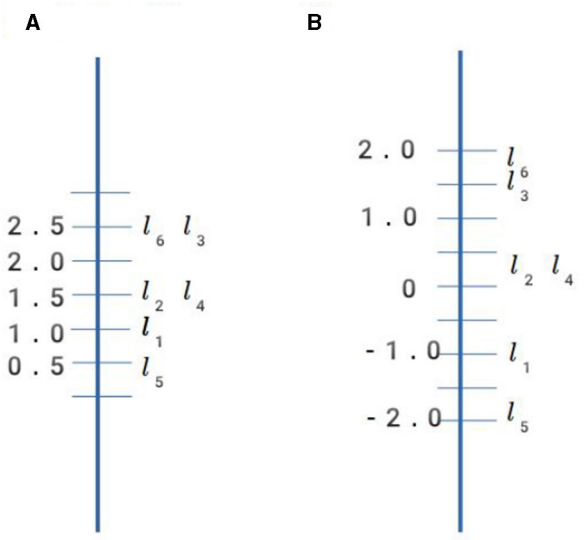



FIGURE 1
IRT-derived proficiency measures. **(A)** Sum score. **(B)** Latent variable score.

the observed data can be characterized as a collection of partially-ordered nodes: a network of "betterness" relations between nodes. In this data model, shown in Figure 2, each node is a *type of performance* on the assessment.

Each type of performance is defined by a collection of *attributes*, that *characterize* it; or (dually) by a collection of *learners*, who *demonstrate* it. The boxes in Figure 2 are the different types of performances on the test. The best performance is at the top of the diagram, and the worst performance at the bottom. Attributes, and learners, may belong to nodes to a *fuzzy degree*. Thus learner 5 belongs to (demonstrates) the lowest type of performance completely (to degree 1). Learners 2, 3, 4, and 6 all demonstrate the highest type of performance to degree 0.5.

*Better* types of performance are characterized by showing *more* attributes (and, dually, are demonstrated by *fewer* learners) than worse types of performance. An arrow from a box $A$ to a box $B$ means that $B$ is a better performance than $A$ (and by extension better than any performance $C$ such that there is a connected path from $C$ to $A$). If there is no path between two types of performance, then they are not comparable. Locating a learner (measuring their proficiency), with respect to this data model for the construct which the three-item test aims to assess, then means finding the "highest"

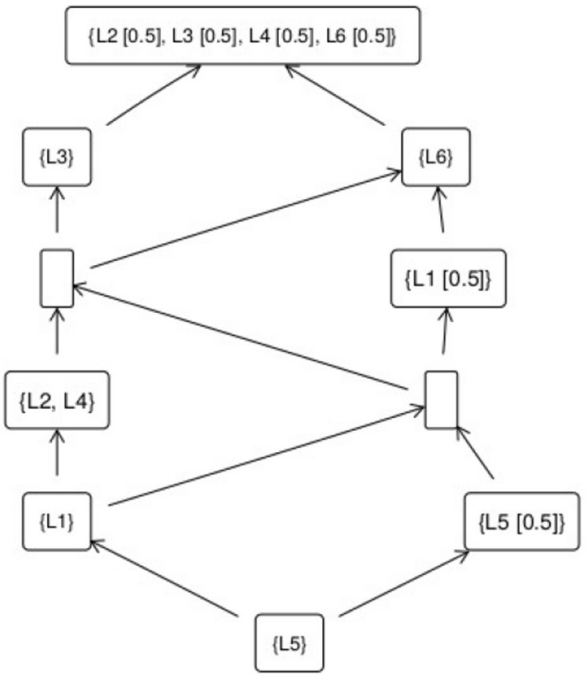



FIGURE 2
Fuzzy concept lattice for assessment data.

node that they belong to in the network. This intuitive description is made more precise in the following section.

## 4.2 Formal concept analysis and proficiency measurement

Formal concept analysis (Ganter and Wille, 1999; Carpineto and Romano, 2004) is an important development of mathematical order theory that has been applied extensively to fields such as linguistics, political science, information sciences, medicine, and genetics. A recent application (Bradley et al., 2024) is to elucidating the mathematical representation of structure in large language models such as ChatGPT, discussed briefly below in Section 6. It can be thought of as a way of making explicit the information structure that is implicit in a matrix—such as that in Table 1—which relates objects to attributes (or learners to test items). It provides methods to extract the concepts and implications that can be deduced from such data, and introduces a logic to reason and infer new knowledge.

Consider first the case of measuring proficiency in a domain by administering an $n$-item test to $m$ learners, where each item is dichotomously scored, i.e., for each learner $l$ and item $i$, it is either the case that $l$ answered $i$ correctly, or that $l$ did not answer $i$ correctly. Given a subset of learners $L_1 := \{l_1, \ldots, l_k\}$, let $I_1 := \{i_1, \ldots, i_j\}$ be precisely those items that all learners in $L_1$ got correct. Then the pair $(L_1, I_1)$ is an instance of a *formal concept* present in the data. $L_1$ is called the *extent* of the concept, and $I_1$ is called its *intent*. We can equally well start with a subset $I_2 := \{i_1, \ldots, i_p\}$ of items, and then form the concept $(L_2, I_2)$, where $L_2$ is precisely the set of learners who got all items in $I_2$ correct.

The collection of all formal concepts extracted from a matrix or data table simply restates the information present by virtue of the way the data is structured due to the choice of attributes (test item responses, in this example), and the ordered valuations chosen for attributes (just the two categories $1 \geq 0$ in this case). However, it makes this structure more apparent (and graphically representable, as in Figure 1) because concepts are (partially) *ordered* via the set-theoretic notion of inclusion. A concept $(L_1, I_1)$ is *more general* than a concept $(L_2, I_2)$ if $L_1 \supseteq L_2$ (or equivalently, if $I_1 \subseteq I_2$). The most general concept is the one that has the largest extent (and smallest intent). In test performance terms, the most general concept corresponds to the bottom, or worst, performance: because every other performance has a larger intent (entails more correct items). Similarly, the least general concept (with the smallest extent and largest intent) corresponds to the top, or best, level of performance.[7]

We can think of formal concepts as different ways of performing on the test (i.e., different ways of exhibiting proficiency in the subject domain). Each type of performance—or exhibition of proficiency—can be described *extensively*, by showing the learners who demonstrated it. Or it can be described *intensively*, by showing the item-profiles that characterized it. These two modes of presentation correspond to different ways of training "measuring instruments" (traditionally, human judges; more recently machine-learning methods such as neural nets) to recognize what good/bad performance (high/low proficiency) looks like. One can either give *examples* of a certain kind of performance, until an assessor can correctly classify new instances, or one can give *descriptions* of that kind of performance (in this case, the relevant profile of item responses), to enable new instances to be classified (measured) correctly.[8]

For a small educational measurement procedure of this kind (small in terms of the number of items/tasks/relevant attributes on which data is collected, as well as small in terms of the number of subjects to which it is administered), the qualitative equivalent of a quantitative score is a learner's location in the concept lattice: the highest concept, in the partial order, to whose extent they belong. This level of proficiency is described, not as a numerical "amount" (location on a line), but rather by the intent of the relevant concept: the actual items they mastered (or, more generally, the construct-relevant attributes

7 Normally concept lattices are drawn as so-called Hasse diagrams with the least general concept at the bottom, and the most general concept at the top. An arrow is drawn upwards from concept $A$ to concept $B$ if $B$ is more general than $A$. In the educational assessment context, we naturally regard the best performance as the *top* concept, which means we need to reverse the usual ordering (in mathematical terms, we use the *dual* lattice). This is done throughout this paper, for example in Figure 2, where the worst level of proficiency (exhibited, to degree 0.5, by learner $l_5$) is at the bottom of the diagram, and the best level (exhibited by learners $l_2, l_3, l_4$, and $l_6$, also to degree 0.5) is at the top.

8 As Weyl (1952, p. 8) noted, "For measurement the distinction is essential between the 'giving' of an object through individual exhibition on the one side, in conceptual ways on the other".

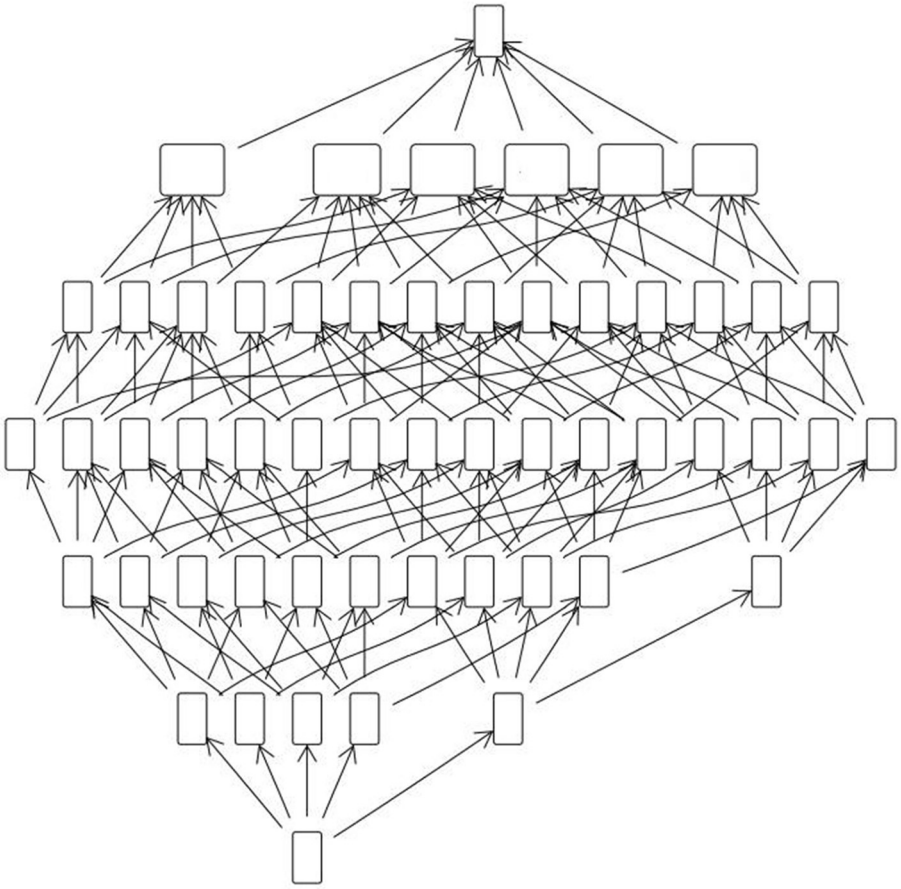


FIGURE 3
Concept lattice for a 5-item test with 100 learners.

their performance demonstrated). For larger (more realistically sized) assessments, the concept-lattice data-model becomes too granular, as shown in Section 5, and we develop a notion of "prototypical" kinds of performances at a manageable number of levels, such that each learner's level, or state, of proficiearency can be described approximately in terms of its qualitatively closest prototype.

Before moving on to that discussion, it is necessary to consider the question of the fuzziness of the criteria that structure data models in many educational measurement procedures.

## 4.3 Truth degrees and fuzzy concepts

### 4.3.1 Assessment results as truth degrees

Table 1 illustrates a situation that often obtains in educational assessment. Learners are given tasks, such as questions on a test, and they may be successful in engaging with them *to a certain degree*. The outcome of a learner's interaction with an item is not necessarily captured by the crisp dichotomy of {correct, incorrect}.

The usual way of dealing with this in psychometric models is to model response categories for polytomous items as a sequence of threshold points on a latent quantitative continuum. A learner's response is in a higher category if it results from their proficiency-state being higher than, but not otherwise different from, a learner whose response is in a lower category. Differences in proficiency must be conceived of as differences in degree, not in kind. Yet as Michell (2012, p. 265) notes, in the context of mathematics tests, "the differences between cognitive resources needed to solve easy and moderately difficult items will not be the same as the differences between resources needed to solve moderately difficult and very difficult mathematics items. This observation suggests that abilities are composed of ordered hierarchies of cognitive resources, the differences between which are heterogeneous."

An alternative approach is to start by the viewing the dichotomous situation as providing information about learners' performances in the form of *propositions* of the form "learner $l$ answered item $i$ correctly".[9] This proposition is true just in case the $(l, i)$ entry in the data table arising from the assessment is 1. So we can think of the entries in the table as truth values (with 0 meaning false and 1 meaning true).

9 As Michell (2009) observes, "Tabulated numbers are shorthand for a set of propositions that tell where the numbers came from. Furthermore, deductions from a data set are inferences from these propositions."

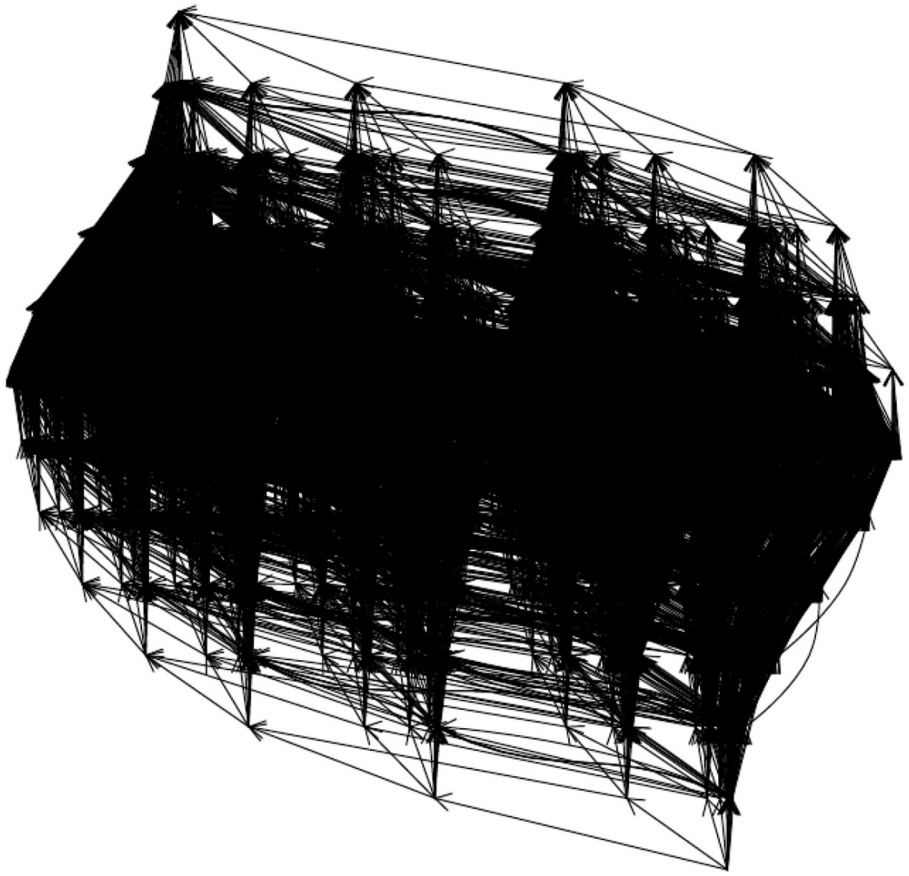


FIGURE 4
Concept lattice for a 12 item test with 200 learners.

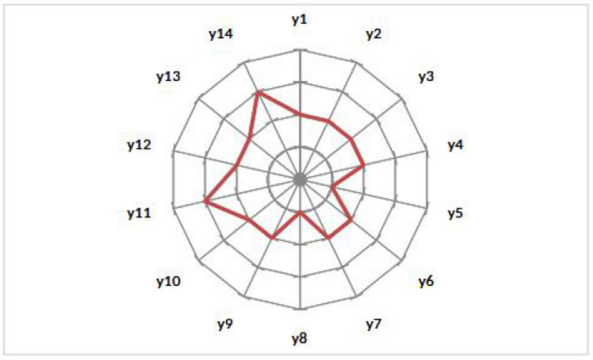
It has long been recognized that, in situations in which there is inherent fuzziness, vagueness, or semantic uncertainty in concepts, bivalent logics, in which the only possible truth values for a proposition are {`false`, `true`} can be unduly restrictive (see e.g., Goguen, 1969; Goertz, 2006; Bělohlávek et al., 2017). *Fuzzy logic* (Hajek, 1998; Bělohlávek et al., 2017) allows propositions to have truth values drawn from ordered sets of *truth degrees*, that can be more extensive than {`false`, `true`}.

Thus we can view the example in Table 1 as providing information about propositions with three truth-degrees, that we could label $\{0, \frac{1}{2}, 1\}$, or {`false`, `partially-true`, `true`}. For example, it is false that learner $l_1$ demonstrated attribute $i_1$ (or we could say, she demonstrated it to degree 0), and it is partially-true that she demonstrated attribute $i_2$ (she demonstrated it to degree $\frac{1}{2}$).

When the outcomes of educational measurement procedures are not completely and crisply dichotomous with respect to all the construct-relevant attributes about which information is collected, the concept lattice for the resulting matrix of fuzzy truth values is itself fuzzy. Objects and attributes belong to concepts with degrees of truth, rather than crisply. In the concept lattice in Figure 2, the label "0.5" after a learner-identifier means that learner belongs to the concept (i.e., has demonstrated that type or level of performance) to degree $\frac{1}{2}$).

Although a discussion of the concept of "measurement error" in psychological testing and educational assessment would take us beyond the scope of this paper, it may be worth clarifying, for the avoidance of doubt, that the application of fuzzy logic in this context is not simply an alternative to using probability theory. Probability is a tool that can be used to study (epistemic) *uncertainty* (the lack of precision that arises from incomplete or poor information), whereas fuzzy logic is a tool that can be used to study (ontological) *vagueness* (the inherent fuzziness, or necessary inexactness, of concepts like "proficiency" in a certain domain). Erwin Schrödinger, when considering what the development of quantum mechanics meant for the measurement of physical phenomena, distinguished these two facets when he noted (Trimmer, 1980; p. 328) that "There is a difference between a shaky or out-of-focus photograph and a snapshot of clouds and fog banks".

The statement "Mary has a fairly good understanding of physics" is vague but certain, whereas "Mary will pass the physics test tomorrow" is precise but uncertain. Working with propositions such as the former (i.e., deploying what Goguen, 1969 calls a "logic of inexact concepts") is core to educational assessment,




Intent of largest factor: 'a medium/reasonable performance across the content and assessment objectives, but characterised by good communication skills on the longer essay items, and less so on the short answers'

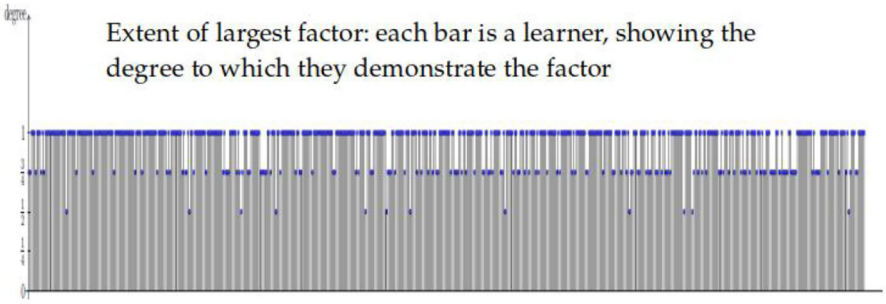



FIGURE 5
Factor representation for fuzzy data.

because of the contestable and intersubjective nature of educational constructs, discussed further in Section 7.2.

### 4.3.2 Truth degrees and quantities

Buntins et al. (2016) apply fuzzy logic to psychological tests in a somewhat different way to that proposed here. They take the view that scores obtained from a test should not "refer to latent variables but to the truth value of the expression 'person *j* has construct *i*", where a *construct* is defined by a collection of relevant *attributes*, each of which may be *possessed* by a test-taker to a certain degree, and each of which may be *relevant* for the construct to a certain degree. Modeling truth degrees as real-valued quantities in the interval [0,1], they present an algorithm for aggregating them across attributes to arrive at an overall score for each learner: the truth value of the proposition "this learner has the construct". They are careful to distinguish the semantic vagueness of a construct definition (recognized in the use of fuzzy truth values) from the idea of "measurement error".

Buntins et al. claim that this approach "neither relies on latent variables nor on the concept of [quantitative] measurement". However, they do state it is arguable that "although there is no measurement theory involved in the ... formalism, the application to actual test behavior does presume item answers to be assessed on an interval scale level", because "test answers have to be real numbers between 0 and 1, reflecting the subjective truth-values of the corresponding attributes for the tested person … However, these only refer to the item level and do not extend to theories about latent variables."

In fact truth degrees do *not* have to be real numbers between 0 and 1. What is required is that they have a way of being compared with each other—that is, an order structure (which could be a partial order)—and way of being combined with each other. In general these requirements are met by taking them to have the mathematical structure of a so-called complete residuated lattice (Hajek, 1998). Further work on conceptualizing truth degrees—and especially what that means for empirically eliciting them—is important, as touched on in Section 7, but beyond the scope of this paper.

Buntins et al. see their approach "not as opposed to psychometric theory but tr[ying] to complement it with an alternative way to conceptualize psychological tests". By contrast, the approach presented in this paper is suggested not as an alternative to, but an extension of, psychometric theory: one in which quantitative measurement forms an important, but special, case of a more general measurement framework.

### 4.3.3 Fuzzy relational systems

In summary, the argument in this section is that in general, educational assessment procedures that aim to measure constructs such as proficiency, ability, or competence in a fuzzily-defined domain, generate *fuzzy relational systems*: matrices of truth-values for propositions of the form "learner *l* has demonstrated construct-relevant attribute *i*". As data models, these are equivalent to fuzzy concept lattices: partially-ordered hierarchies, or networks, of types of performance on the assessment, that are discriminable with respect to these construct-relevant attributes. The next section

considers whether these data models can provide insight for realistically-sized assessments.

# 5 Practicalities of educational assessment with non-quantitative data models

## 5.1 Granularity of data models

An issue with data models of the kind discussed in the previous section is that their combinatorial complexity increases geometrically with the numbers of learners and construct-relevant attributes of performance (or test items) involved. Figures 3, 4, for instance, show the concept lattices for subsets of outcomes of a physics test.[10] with increasing numbers of learners and attributes. Clearly the information here is too granular to be useful, and we need to simplify or "smooth" it in some way.

For quantitative data models, where learners' test responses as thought of as vectors in $n$-dimensional Euclidean space, the analogous granuarity-reduction is often performed using latent variable models that aim to find a $k$-dimensional subspace with $k < n$ (often a one-dimensional subspace, i.e., a line) that is oriented in such a way as most closely to approximate the direction of most of the variation between the positions of these points (possibly subject to some other constraints as well, for certain factor-analytic models: see Bartholemew et al., 2008). Each learner's latent-variable score is then the projection of the vector that represents their test performance onto this subspace. Calculating these scores entails factorizing the (transpose of the) matrix $Z$ of normalized test scores. If there are $m$ learners and $n$ test items, then the $n \times m$ item-by-learner matrix $Z^T$ is factorized into the product of a $n \times k$ item-by-factor matrix $L$ and a $k \times m$ factor-by-student matrix $F$, plus some error: $Z^T \approx LF$. Then using standard results in linear algebra, it can be shown (e.g., Reyment and Jöreskog, 1993) that the factors are the eigenvectors of the covariance matrix $ZZ^T$.

## 5.2 Factorizing qualitative matrices

Bělohlávek (2012) studied the question of factorizing a matrix of fuzzy truth values. Now the matrix product is no longer defined in terms of operations on quantities, but rather in terms of operations on truth values.[11] Let $M$ be an $m \times n$ matrix arising from an educational measurement procedure conceptualized as in Section 4.3, so that $M_{ij}$ is the degree to which learner $i$ displays attribute $j$. By analogy with the quantitative case, consider an approximate factorization of $M$ into a $m \times k$ learner-by-factor matrix $A$ and a $k \times n$ factor-by-attribute matrix $B$, i.e., $M \approx A \circ B$. The key theorem in this case, due Bělohlávek (2012), is that *the factors are particular formal concepts* from the concept lattice for $M$. That is, "picking out key concepts" (particular types of learners' responses to the assessment) is equivalent to "logically factorizing" the matrix of truth-degrees that is the outcome of the measurement procedure.

The factors are the (extents and intents) of specific concepts in the concept lattice for $M$. The intuition is that, with $M_{ij} = A_{ip} \circ B_{pj}$:

- $A_{ip}$ is the degree to which learner $i$ is an example of (in the extent of) factor $p$;
- $B_{pj}$ is the degree to which attribute $j$ is one of the manifestations of (in the intent of) factor $p$;
- $M = A \circ B$ means: learner $i$ displays attribute $j$ if and only if there is a factor (formal concept) $p$ such that $i$ is an example of $p$ (or $p$ applies to $i$); and $j$ is one of the particular manifestations of $p$.

Thus, the qualitative analog of projecting a Euclidean space onto a lower-dimensional subspace consists in picking out certain points in a partially ordered set. Specific formal concepts are selected, similarly to the way in which specific vectors—the eigenvectors of the covariance matrix—are selected when learners are scored on quantitative latent variables. The analogs of scores on a latent variable are the degrees to which learners' performances "display" or "participate in" or "reflect" these specific concepts, which may be thought of as *prototype* or *standards of performance* on the construct. They have the advantage, over hypothesized latent variables whose values are abstracted from observed data, that they are directly expressible in terms of the construct-relevant attributes—that is, in terms of the features of learner's responses to assessment tasks that are taken to be important in a "theory" of "what (good) performance means", for the educational construct in question. They can be described both by means of their extent (the collection of actual learners' performances exemplifying the concept/standard in question), and by means of their intent [the collection of (fuzzy) attributes that characterizes the standard in question].

## 5.3 Measures and meanings: comparing quantitative and qualitative approaches

Bartl et al. (2018) examined this qualitative factor analytic approach to educational assessment data, with the aims of exploring its applicability in practice, and its application to the study of the construct validity of an examination: the degree to which students' responses, assessed as being at a particular level, matched the intentions of the assessment designers in terms of the qualitative performance standard intended to broadly characterize responses at that level. This is the kind of question that is difficult to study using traditional quantitative methods.

10 Part of paper 1 of the AQA A level physics examination taken in 2018. Unusually for an A level assessment, the items here are all dichotomous (multiple-choice questions). The lattices would be even larger if the items admitted fuzzy valuations.

11 The product of two real-valued matrices $A$ and $B$ is defined by setting its $(i,j)$ entry $(AB)_{ij}$ to the inner product of row $i$ of $A$ with column $j$ of $B$: i.e., $(AB)_{ij} := \sum_{p=1}^{k} A_{ip}B_{pk}$. When the matrix entries are truth values, they are elements of a type of lattice that is equipped with an operation $\otimes$ to combine values. In this case the matrix product $A \circ B$ is defined as $(A \circ B)_{ij} := \bigvee_{p=1}^{k} A_{ip} \otimes B_{pk}$, where $\bigvee$ is the supremum over the indicated set (see Appendix).

The technical issues involved (for example how to determine the coverage and number of factors that broadly explain the data—analogous to a scree plot in quantitative principal components analysis) will not be rehearsed here. See Bartl et al. (2018) for computational details. For a deeper theoretical treatment of the relationship between eigenvectors (of quantitative covariance matrices) and formal concepts (of qualitative matrices of truth values), see Bradley (2020). The key point is that this approach allows drawing out key features associated with responses assigned to a particular level, by the assessment procedure, and an appraisal of the degree to which each learner's performance on the examination embodies or matches those features. Indeed, it "explained" the data (in terms of proportion of data covered or variance explained) as well as standard principal components analysis, but generated factors exemplifying attributes of performance that seemed to be more easily interpretable.

Figure 5 shows an example of this, for the educational measurement data studied by Bartl et al. (2018), in which learners were assessed on 14 fuzzy attributes $\{y_1, \ldots, y_{14}\}$, each of which reflected an aspect of the construct, in this case proficiency in the specific subject of "A level Government and Politics". Each of the attributes corresponds to demonstrating specific types of knowledge and understanding, in accordance with the examiners' agreed understanding of what better/worse proficiency means in this domain. Hence the intent of any given concept can be interpreted by users of the assessment as a description of broadly what that level of proficiency means (and likewise the extent of the concept can be interpreted as an indication of the degree to which each learner has demonstrated that level of proficiency).

The question of the interpretability or explainability of the results of educational measurement procedures—whether those results are numerical scores, or broader grades or levels—is particularly important for high-stakes assessments such as those that underwrite school-leaving qualifications. For learners, clarity about *why* their response to an assessment merited their being characterized as demonstrating a certain level of proficiency is arguably required for reasons of natural justice. For teachers, understanding qualitatively what their students did well, and what they would have to do better to demonstrate more proficiency in a subject domain, is clearly valuable as an input into their future pedagogical practice. Bartl et al. (2018, p. 204) concluded that their approach to qualitative factor analysis yielded "naturally intepretable factors from data which are easy to understand", but that more research is needed both on technical implementation and on the views of learners and teachers.

## 5.4 Other order-theoretic approaches to educational assessment

In the 1940s Louis Guttman began to develop an approach to psychological measurement (e.g., Guttman, 1944) that led him to think of it as a structural theory (Guttman, 1971), rather than as a process of quantifying amounts of latent traits, and to the development of *facet theory* and *partial order scalogram analysis* (Shye and Elizur, 1994). In the 1980s, Doignon and Falmagne (1999) developed *knowledge space theory*, later evolved into a theory of learning spaces, in which assessment constructs are represented as partially-ordered sets.

Applications of facet theory and knowledge space theory (including related approaches such as Tatsuoka, 2009's *rules space* and Leighton and Gierl, 2007's *cognitive diagnostic models*) normally assume or overlay quantitative latent variable models, to account for "underlying" proficiencies or competencies that determine a learner's progression through such partially-ordered outcome spaces.

However, from the mid 1990s onwards, there has been a strand of research investigating how to extend knowledge space theory to incorporate a focus on skills and competence, leading to the development of *competence-based knowledge space theory* (see e.g., Stefanutti and de Chiusole, 2017). Here, a learner's proficiency or competence is itself conceptualized as a partially-ordered space, rather than a quantity. Ganter and Glodeanu (2014) and Ganter et al. (2017) suggested that formal concept analysis could be applied to study competence-based knowledge space theory, and this is now starting to be done.

For example, Huang et al. (2023) consider how to transform maps from competence-states to "knowledge-states" (types of demonstrated performances) into formal contexts, and hence to represent them as concept lattices. Each node in the lattice then embodies a knowledge-state and a competence-state as its extent and its intent, respectively. This is clearly analogous to the approach set out in Section 4 above.

A very clear application of these methods is to formative, adaptive, assessment and learning systems, where, for instance, they provide an alternative to traditional IRT-based adaptive tests that is more grounded in a theory of learning.

To date there has been less attention to examining summative assessment, and what is often called "educational measurement", from this perspective. Yet, as argued above, application of non-quantitative approaches needs to be investigated here too, since the pragmatic "as if" approach to routine application of latent variable models is not always justifiable.

# 6 Connections to artificial intelligence

A final reason why it is imperative to pursue research in this area is the rapidly growing application of machine-learning methods, and generative artificial intelligence in particular, in educational contexts. For example, Li et al. (2023) report on using the large language model ChatGPT to score students' responses to (essay style) examinations, and to provide rationales for the scores awarded.

Because the outputs of generative AI applications using large language models are no more than statistically plausible sequences of words, albeit expressed in well-formed natural language, their validity, fairness and reliability is hard to establish theoretically. That is because they are produced using so-called *subsymbolic* approaches to AI (see e.g., Sudmann et al., 2023), such as deep neural nets, rather than *symbolic* methods that aim to use forms of explicit logical inference to arrive at results: analogously to reasoning about a learner's response to a task with reference to criteria for betterness that define the kind of proficiency one intends to measure by administering the task.

An interesting angle opened up by the qualitative measurement approach described above is the possibility of combining formal concept analysis with neural networks to enhance the explainability of, for example, scores derived from applying a classifier based on a large language model to learners' performances on an examination.

Some initial work in this area has been done by Hirth and Hanika (2022) and Marquer (2020), among others. This kind of analysis could complement quantitative approaches to explaining marks or scores awarded to learners' responses, such as dimension-reduction of the high-dimensional vector space that the language model uses to represent linguistic artifacts—such as learners' responses to assessment tasks—as numerical vectors. In fact, Bradley et al. (2024) have recently shown that there is a relationship between quantitative techniques based on linear algebra, such as latent semantic analysis, and formal concept analysis, such that the latter can be seen as a more general form of the former. They have applied formal concept analysis to elucidating how semantics appears to arise from syntax, and to study the structure of semantics, when large language models are used to produce outputs from qualitative data.

Clearly, the practice of educational (and psychological) measurement is changing as technology changes. Tasks can be administered digitally; the widespread availability of devices with reasonable processing power means the possibilities for task design are much more open than they were a decade ago, and they will continue to evolve. The data that is gathered about learners, given their responses to these tasks, can be more unstructured than category-labels or scores: it may be text, audio, or video, and/or representations of such data for example in a vector-space language model. To the extent that human assessors form part of measurement procedures, for example to apply scoring rubrics, they may be partially or wholly replaced by AI.

What remains fundamental, however, is the need to base these measurement procedures in a theory of what defines or constitutes better or worse proficiency, in the domain of interest, and hence what substantive and semantic content is entailed in statements such as "this learner got a score of 137", or "this learner has 1.07 logits of proficiency"; or "this learner has demonstrated three of the four prototypical aspects of proficiency that define a "grade B standard", or whatever — what it means to locate them, via a measurement, at a certain position in a (quantitative or other) space.

# 7 Discussion

## 7.1 Qualitative educational assessment is possible in principle, and includes quantitative measurement as a special case

This paper has argued that it is not warranted to assume the phenomena studied in psychometrics, and in educational measurement in particular, are necessarily appropriately conceptualized as quantities. In cases where an assumption of quantitative structure *is* appropriate, then measuring an instance of such a phenomenon means locating it at a point on the real continuum. In cases where the assumption is not appropriate, the idea of measurement becomes, more generally, locating the measurand in a suitable logical space, that is defined in a way that is relevant for the phenomenon.

When the measurand is quantitative and the logical space is the real numbers, the usual methods of psychometric analysis for estimating latent parameters can be deployed. But, *contra* Thurstone (1928), the paper has argued that it is not necessary to "force" theoretically well-supported constructs into a more reductive quantitative form if that is not appropriate. Hence the argument of this paper is not that psychometrics should be replaced, but that its repertoire of measurement approaches should be widened to cope with measurands that are intrinsically non-quantitative in nature.

The paper suggests that the outcomes of educational measurement procedures can be thought of, in general, as fuzzy relational systems; and that fuzzy formal concept analysis is an appropriate tool to describe data models for the measurands they aim to locate. These models instantiate the "betterness" relation for the measurand: they model the notion of "what good performance looks like". Such an account or understanding is prior to, and necessary for, an understanding or agreement as to "what being (more or less) proficient" means, in an educational domain. It forms the theory of the construct (one might say, the theory of *value* for the construct, and hence a foundation for evaluation of construct *validity*).

## 7.2 Educational constructs are contestable, intersubjective, temporally-located phenomena

These theories of constructs such as proficiency or competence in a domain are necessarily contestable, intersubjectively constructed, and liable to change over time. Intersubjectivity (Chandler and Munday, 2011) refers to the mutual construction of relationships through shared subjectivity. Things and their meanings are intersubjective, within a given community, to the extent that the members of the community share common understandings of them. Thus, the community that constitutes the competent authority for defining an educational construct decides what particular knowledge, skills, and understanding it will encompass, and what will count as better or worse configurations of these aspects as possible ways of being proficient in the domain in question. Thus, for instance, the job of someone marking responses to an examination that is designed to measure that construct is to apply the mutually constructed and agreed standard consistently to each response she marks (irrespective of whether she personally agrees that it is the "right" standard).

We do not have to think of data models that encode these intersubjective constructions as (more or less accurate) representations of some objective or underlying "true" account of the measurand in question. As van Fraassen (2008, p. 260) notes, "in a context in which a given model is *someone's* representation of a phenomenon, there is **for that person** no difference between the question *whether a theory fits that representation* and the question *whether that theory fits the phenomenon*."

### 7.3 More research is needed on using partial orders in practice, on linking different assessments of the same construct, and on fuzzy valuations

Section 4 argued that in general the data models for measurands such as proficiency in an educational domain are partial orders. This perhaps goes against a relatively strongly ingrained concept of educational assessment as synonymous with *ranking* (e.g., Holmes et al., 2017). Yet in many cases, once a theory of (betterness for) a construct has been settled, rankings are neither necessary nor needed. Two learners' proficiency values may simply be qualitatively different (non-comparable). For instance in Figure 2, this is the case for learners 3 and 6. But both learners 3 and 6 have performed better than learner 1. So if learner 1's performance was sufficient to merit a "pass" grade, let us say (or was picked out as a "pass" grade prototype), then we know that learners 3 and 6 are also sufficiently proficient to be awarded a pass, even though it is not meaningful to say that their actual demonstrated proficiencies were the same, or that either one is more or less proficient than the other. More work is needed on the scope for using visualizations such as concept lattices to help educational assessment designers and teachers engage with and interrogate the outcomes of educational measurement procedures (see, for a start, Bedek and Albert, 2015).

A common application of quantitative latent variable models is to *equating* or *linking* different forms of tests of learners' proficiency in a certain domain. Typically, equating studies are designed to answer questions like "what score on form $X$ of a test is equivalent to (represents the same level of proficiency as) a given score on form $Y$ of the test?". In practical applications in many educational contexts however, such as grading students' responses to school-leaving examinations (Newton et al., 2007), one is not so much interested in constructing a monotone map from scores on $X$ to scores on $Y$, as in ensuring that the levels or kinds of proficiency demonstrated by students graded, say, A, on this year's examination, are "equivalent", or "of a comparable standard" to the type of proficiency demonstrated by students graded A on last year's examination.

An area for further research is how to implement such comparability studies in the fuzzy-relational approach to educational assessment proposed in this paper. For example one could take the students graded A on each of the two forms of an assessment, and examine the intents of the formal concepts that form their largest factors (cover an appreciable proportion of the data, in the terms of Bartl et al., 2018). Are these sufficiently similar to count as equivalent demonstrations of proficiency, and what criteria should be applied to appraise similarity?

A deeper question is how the truth degrees that summarize each learner's demonstration of each construct-relevant attribute are determined. In some cases this is straightforward in practice (e.g., for dichotomously-classified test items such as multiple-choice questions); but when judges are needed as part of the measurement procedure, different judges may give different truth values, so what counts as a reasonable or acceptable value? A full account of this aspect of qualitative valuation may need to draw on *rough fuzzy logic* (Dubois and Prade, 1990; Bazan et al., 2006), itself an active area of research in machine learning. Certainly more research is needed here.

Having said that, there is strong support for connecting fuzzy relational structures to cognitive theories of concept formation, when exploring the question of how experts—and these days, AIs—learn to categorize (value) responses to tasks, given some prototypical exemplars: see for example Bělohlávek and Klir (2011).

The outcomes of educational measurement procedures are ultimately underpinned by value judgements about exactly what to assess and how to assess it. As Wiliam (2017, p. 312) puts it: "whereas those focusing on psychological assessment tend to ask, 'Is this correct?', those designing educational assessment have to ask, 'Is this good?'". So questions about how to use mathematical methods in these contexts, in a way that leverages their power, but is not unduly reductive, will no doubt always be debated. It is hoped this paper makes a helpful contribution to that debate.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

AS: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Acknowledgments

AS would like to thank his Oxford DPhil supervisor Associate Professor Joshua McGrane (Graduate School of Education, University of Melbourne) for helpful comments on an early draft and general discussion and constructive criticism of some of the ideas presented here. He would like to thank his Oxford DPhil supervisor Professor Jenni Ingram (Department of Education, University of Oxford) for overall support with this research. He also thanks the editor and reviewers for their very helpful comments on an earlier draft of this paper.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1399317/full#supplementary-material

## References


Andrich, D., and Marais, I. (2019). *A Course in Rasch Measurement Theory*. Singapore: Springer.

Bartholemew, D., Steele, F., Moustaki, I., and Galbraith, J. (2008). *Analysis of Multivariate Social Science Data*. Boca Raton, FL: CRC Press.

Bartl, E., Bělohlávek, R., and Scharaschkin, A. (2018). "Toward factor analysis of educational data," in *Proceedings of the 14th International Conference on Concept Lattices and their Applications*, eds. D. Ignatov, and L. Nourine (Olomouc), 191–206.

Bazan, J., Skowron, A., and Swiniarski, R. (2006). "Rough sets and vague concept approximation: from sample approximation to adaptive learning," in *Transactions on Rough Sets V: Lecture Notes in Computer Science 4100*, eds. J. Peters, and A. Skowron (Berlin: Springer), 39–62.

Bedek, M., and Albert, D. (2015). "Applying formal concept analysis to visualise classroom performance," in *Proceedings of the 11th International Conference on Knowledge Management*, eds. T. Watanabe, and K. Seta (Osaka).

Bělohlávek, R. (2012). Optimal decomposition of matrices with entries from residuated lattices. *J. Logic Comp*. 22, 1405–1425. doi: 10.1093/logcom/exr023

Bělohlávek, R., Dauben, J., and Klir, G. (2017). *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford: Oxford University Press.

Bělohlávek, R. and Klir, G. (eds.). (2011). *Concepts and Fuzzy Logic*. Cambridge, MA: The MIT Press.

Bradley, T.-D. (2020). *At the Interface of Algebra and Statistics* (PhD thesis). New York, NY: City University of New York.

Bradley, T.-D., Gastaldi, J., and Terilla, J. (2024). The structure of meaning in language: parallel narratives in linear algebra and category theory. *Not. Am. Math. Soc*. 71, 174–185. doi: 10.1090/noti2868

Buntins, M., Buntins, K., and Eggert, F. (2016). Psychological tests from a (fuzzy-)logical point of view. *Qual. Quant*. 50, 2395–2416. doi: 10.1007/s11135-015-0268-z

Carpineto, C., and Romano, G. (2004). *Concept Data Analysis: Theory and Applications*. Chichester: Wiley.

Chandler, D., and Munday, R. (2011). *A Dictionary of Media and Communication*. Oxford: Oxford University Press.

Cho, E. (2023). Interchangeability between factor analysis, logistic irt, and normal ogive irt. *Front. Psychol*. 14:1267219. doi: 10.3389/fpsyg.2023.1267219

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxf. Rev. Educ*. 34, 609–636. doi: 10.1080/03054980801970312

Doignon, J.-P., and Falmagne, J.-C. (1999). *Knowledge Spaces*. Berlin: Springer.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika* 79, 1–19. doi: 10.1007/s11336-013-9342-4

Dubois, D., and Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *Int. J. Gen. Syst*. 17, 191–209. doi: 10.1080/03081079008935107

Ganter, B., Bedek, M., Heller, J., and Suck, R. (2017). "An invitation to knowledge space theory," in *Formal Concept Analysis: 14th International Conference, ICFCA 2017* (Rennes: Springer), 3–19.

Ganter, B., and Glodeanu, C. (2014). "Factors and skills," in *Formal Concept Analysis: 12th International Conference, ICFCA 2014* (Cluj-Napoca: Springer), 173–187.

Ganter, B., and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer.

Goertz, G. (2006). *Social Science Concepts*. Princeton, NJ: Princeton University Press.

Goguen, J. (1969). The logic of inexact concepts. *Synthese* 19, 325–373. doi: 10.1007/BF00485654

Guttman, L. (1944). A basis for scaling qualitative data. *Am. Sociol. Rev*. 9, 139–150. doi: 10.2307/2086306

Guttman, L. (1971). Measurement as structural theory. *Psychometrika* 36, 329–347. doi: 10.1007/BF02291362

Hajek, P. (1998). *Metamathematics of Fuzzy Logic*. Dordrecht: Kluwer.

Heene, M. (2013). Additive conjoint measurement and the resistance towards falsifiability in psychology. *Front. Psychol*. 4:246. doi: 10.3389/fpsyg.2013.00246

Heilmann, C. (2015). A new interpretation of the representational theory of measurement. *Philos. Sci*. 82, 787–797. doi: 10.1086/683280

Hirth, J., and Hanika, T. (2022). Formal conceptual views in neural networks. *arXiv [Preprint]*. doi: 10.48550/arXiv.2209.13517

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. Berichteüber die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. *Mathematisch-Physische Klasse* 53, 1–46.

Holmes, S., Black, B., and Morin, C. (2017). *Marking Reliability Studies 2017: Rank Ordering Versus Marking: Which Is More Reliable*? Coventry, UK: Technical Report, Ofqual.

Huang, B., Li, J., Li, Q., Zhou, Y., and Chen, H. (2023). *Competence-Based Knowledge Space Theory From the Perspective of Formal Concept Analysis*. Available at: https://ssrn.com/abstract=4620449 (accessed August 13, 2024).

Kane, M. (2008). The benefits and limits of formality. *Measur. Interdisciplin. Res. Perspect*. 6, 101–108. doi: 10.1080/15366360802035562

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *J. Appl. Meas*. 2, 389–423.

Karabatsos, G. (2018). On Bayesian testing of additive conjoint measurement axioms using synthetic likelihood. *Psychometrika* 83, 321–332. doi: 10.1007/s11336-017-9581-x

Kline, P. (2000). *A Psychometrics Primer*. London: Free Association Press.

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. New York, NY: Academc Press.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance: plausible conjoint systems. *Br. J. Math. Stat. Psychol*. 64, 478–497. doi: 10.1348/2044-8317.002004

Leighton, J., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Li, J., Gui, L., Zhou, Y., West, D., Aloisi, C., and He, Y. (2023). Distilling ChatGPT for explainable automated student answer assessment. *arXiv* [preprint]. doi: 10.18653/v1/2023.findings-emnlp.399

Lord, F., and Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Luce, R., and Narens, L. (1994). "Fifteen problems concerning the representational theory of measurement," in *Patrick Suppes: Scientific Philosopher*, ed. P. Humphries (Dordrecht: Springer), 219–249.

Luce, R., and Tukey, J. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *J. Math. Psychol*. 1, 1–27. doi: 10.1016/0022-2496(64)90015-X

Marquer, E. (2020). *Latticenn: Deep Learning and Formal Concept Analysis* (Master's thesis). Nancy: Université de Lorraine.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measur. Interdiscip. Res. Perspect*. 15, 51–69. doi: 10.1080/15366367.2017.1348108

McGrane, J., and Maul, A. (2020). The human sciences: models and metrological mythology. *Measurement* 152:107346. doi: 10.1016/j.measurement.2019.107346

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. London: Routledge.

Michell, J. (1999). *Measurement in Psychology: Critical History of a Methodological Concept. Ideas in Context* (Cambridge: Cambridge University Press), 53.

Michell, J. (2006). Psychophysics, intensive magnitudes and the psychometricians' fallacy. *Stud. Hist. Philos. Biol. Biomed. Sci*. 17, 414–432. doi: 10.1016/j.shpsc.2006.06.011

Michell, J. (2009). The psychometricians' fallacy: too clever by half. *Br. J. Math. Stat. Psychol.* 62, 41–44. doi: 10.1348/000711007X243582

Michell, J. (2012). The constantly recurring argument: inferring quantity from order. *Theory Psychol.* 22, 255–271. doi: 10.1177/0959354311434656

Michell, J. (2013). Constructs, inferences and mental measurement. *New Ideas Psychol.* 31, 13–21. doi: 10.1016/j.newideapsych.2011.02.004

Michell, J. (2021). Representational measurement theory: is its number up? *Theory Psychol.* 31, 3–23. doi: 10.1177/0959354320930817

Newton, P., Baird, J., Goldstein, H., Patrick, H., and Tymms, P. (eds.). (2007). *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority.

Peres, A. (1995). *Quantum Theory: Concepts and Methods*. Dordrecht: Kluwer.

Pollitt, A., and Ahmed, A. (2008). "Outcome space control and assessment," in *Technical report, Paper for the 9th annual conference of the Association for Educational Assessment–Europe* (Hissar).

Raykov, T., and Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.

Reid, T. (1748 [1849]). "An essay on quantity," in *The Works of Thomas Reid*, ed. W. Hamilton (Maclachlan, Stuart and Co., Edinburgh), 715–719.

Reyment, R., and Jöreskog, K. (1993). *Applied Factor Analysis in the Natural Sciences*. Cambridge: Cambridge University Press.

Scharaschkin, A. (2023). "Measuring educational constructs qualitatively," in *Paper Presented at the Annual Conference of the Association for Educational Assessment Europe* (Malta).

Shye, S. and Elizur, D. (eds.). (1994). *Introduction to Facet Theory*. Thousand Oaks, CA: Sage Publications, Inc.

Stefanutti, L., and de Chiusole, D. (2017). On the assessment of learning in competence based knowledge space theory. *J. Math. Psychol.* 80, 22–32. doi: 10.1016/j.jmp.2017.08.003

Stevens, S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Sudmann, A., Echterhölter, A., Ramsauer, M., Retkowski, F., Schröter, J., and Waibel, A. (eds.). (2023). *Beyond Quantity: Research with Subsymbolic AI*. Bielefeld: transcript Verlag.

Tal, E. (2020). "Measurement in science," in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta (Stanford, CA: Stanford University).

Tatsuoka, K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. Boca Raton, FL: CRC Press.

Thurstone, L. (1927a). A law of comparative judgement. *Psychol. Rev.* 34, 278–286. doi: 10.1037/h0070288

Thurstone, L. (1927b). The method of paired comparisons for social values. *J. Abnorm. Soc. Psychol.* 21, 384–400. doi: 10.1037/h0065439

Thurstone, L. (1928). Attitudes can be measured. *Am. J. Sociol.* 33, 529–554. doi: 10.1086/214483

Trimmer, J. (1980). The present situation in quantum mechanics: A translation of Schrödinger's "cat paradox" paper. *Proc. Am. Philos. Soc.* 124, 323–338.

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theoret. Philos. Psychol.* 41, 58–84. doi: 10.1037/teo0000176

Uher, J. (2022a). Functions of units, scales and quantitative data: fundamental differences in numerical traceability between sciences. *Qual. Quant.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: a rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

van der Linden, W., and Hambleton, K. (eds.). (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer.

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

van Rooij, R. (2011). Measurement and interadjective comparison. *J. Semant.* 28, 335–358. doi: 10.1093/jos/ffq018

von Davier, A., Mislevey, R., and Hao, J. (eds.). (2021). *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment*. Cham: Springer.

Weyl, H. (1952). *Space, Time, Matter*. New York, NY: Dover.

Wiliam, D. (2017). Assessment and learning: a long and winding road. *Assess. Educ.* 24, 309–316. doi: 10.1080/0969594X.2017.1338520

Wolff, J. (2020). *The Metaphysics of Quantities*. Oxford: Oxford University Press.

frontiers | Frontiers in Psychology

TYPE Perspective
PUBLISHED 02 October 2024
DOI 10.3389/fpsyg.2024.1374308



OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
Jan Ketil Arnulf,
BI Norwegian Business School, Norway

*CORRESPONDENCE
Václav Linkov
✉ linkov1@uniba.sk

RECEIVED 21 January 2024
ACCEPTED 13 September 2024
PUBLISHED 02 October 2024

CITATION
Linkov V (2024) Qualitative (pure) mathematics as an alternative to measurement.
*Front. Psychol.* 15:1374308.
doi: 10.3389/fpsyg.2024.1374308



# Qualitative (pure) mathematics as an alternative to measurement

Václav Linkov*

Institute of Applied Psychology, Faculty of Social and Economic Sciences, Comenius University in Bratislava, Bratislava, Slovakia

This paper focuses on the possible usage of qualitative mathematics in psychology. Qualitative mathematics is understood to be equivalent to pure mathematics. First, it is explained that mathematics is a discipline studying patterns in reproducible mental objects. Qualitative mathematics is presented as an alternative to measurement, potentially offering the same level of exactness, clarity, and rigor. This perspective might lead psychologists to explore connections between a phenomenon and any kind of mathematical structure, regardless of whether the structure is quantitative. Usage of (any) mathematical structures might require scholars who are familiar with them. Consequently, changes in mathematics education may also be needed. Introducing non-numerical structures into mathematics education—thereby partially revisiting the New Math Movement—could train individuals more prepared for a creative approach to the use of structures and less inclined to view everything as quantitative.

KEYWORDS

**measurement, qualitative mathematics, quantification, psychology, non-numerical, mathematics education**

## 1 Introduction

There is a long-term debate in psychology about whether all or nearly all psychological phenomena should be quantified and studied using quantitative methods, or if quantification is not a suitable method for the majority of psychological attributes (Toomela, 2008; Uher, 2021; Franz, 2022). Quantification is a topic for psychologists, as quantitative structure is the only part of mathematics typically used in psychology. In this paper I first attempt to explain what qualitative mathematics is. Second, I argue that if qualitative mathematics is to be used, it could serve as an alternative to quantification and measurement, offering the same level of exactness. Third, I argue that this would need change in mathematics education, which primarily focuses on numerical representations in schools.

## 2 Qualitative mathematics

Lee (2013) introduced the term "qualitative mathematics" in the title of his book but did not define it there. This term is not frequently used by mathematicians, and there is no universally accepted definition in mathematics either. However, a commonly accepted meaning among many mathematicians might be that "qualitative mathematics" is synonymous with "pure mathematics." I will elaborate on what this means and the consequences of this interpretation of "qualitative mathematics."

Defining (pure) mathematics is not straightforward. Byers (2017) believes that the best description of mathematics is that it is what mathematicians do. In a modern view, we might

define mathematics as the science of patterns (Devlin, 2012). Another definition might be that offered by Hersh (2014), who states that mathematics involves ideas, concepts, which exists only in the shared consciousness of human beings… is both a science and a "humanity" "(Hersh, 2014, p. 163). He describes it as a discipline studying "mental objects with reproducible properties" (p. 163).

An important characteristic of mathematics is its frequent need to clarify and change terminology until it finds some representation of a problem that allows it to be solved. Ziegler and Loos (2014, p. 1210) state that people are usually not aware that part of mathematics "is a struggle to find and shape the 'right' concepts/definitions and to pose/develop the 'right' questions and problems." This notion is further developed by Schwartz (2006, p. 232): "Mathematics must deal with well-defined situations. Thus, in its relations with science, mathematics depends on an intellectual effort outside of mathematics for the crucial specification of the approximation which mathematics is to take literally." This "well-defined" does not mean that terminology must be precisely defined. When Newton and Leibniz developed calculus, they did not have precisely defined terminology for "continuous," and they relied on intuitive understanding—this term was precisely defined by Bolzano more than 100 years later (Boyer, 1949). Formalizing a problem into precise terminology can be difficult, and some mathematicians believe that the most important part of problem-solving involves unconscious processes (Hadamard, 1945).

Let us consider some relations between mathematics and psychology. William James defines psychological phenomena as "such things as we call feelings, desires, cognitions, reasonings, decisions, and the like" (James, 1950, p. 1). Here we see that both mathematics and psychology study objects that exist only in the human mind. Since Cantor (1895, p. 481) defined a set as "any collection M of definite, well-distinguished objects m of our perception or our thought," psychological phenomena might also be considered as a set (if they are "well-distinguished"). A logical consequence might be to look for similarities between mathematical and psychological objects so that mathematical objects could be used as representations of psychological ones.

If we interpret "qualitative mathematics" to mean "pure mathematics," the counterpart to this is applied mathematics. Both disciplines deal with mathematical objects as their subject matter, but their objectives and approaches differ. Higham (2015, p. 1), in attempting to describe what this difference in objectives and approaches entails, notes that defining it is nearly impossible; hence, he cites the perspectives of several scholars without providing a concrete source. Applied mathematics could be described as "the bridge connecting pure mathematics with science and technology," according to William Prager. Richard Courant offers a deeper insight, stating that "Applied mathematics is not a definable scientific field but a human attitude… [the scientist] must be willing to make compromises regarding rigorous mathematical completeness." The third perspective that Higham includes is from Peter Lax, who remarks that "the applied mathematician must rely on… special solutions, asymptotic descriptions, simplified equations, experimentation both in the laboratory and on the computer." The main difference in objectives is that while pure mathematics focuses on theoretical understanding, applied mathematics is concerned with practical applications in the external world. The difference in approach is that pure mathematics seeks to comprehend why something is valid, whereas applied mathematics is satisfied if it provides reproducible results. Applied mathematics does not concern itself with understanding the underlying reasons, thus it is less reflective of the theoretical aspects.

Understanding that "qualitative mathematics" encompasses all mathematical objects is evident in Lee's (2013) book. One chapter discusses complex dynamical systems, which are systems utilizing nonlinear functions over a quantitative structure. These systems necessitate the measurement of quantitative variables. In Lee's text, the quantitative aspect is merely one instance of the qualitative. What characterizes mathematics as qualitative is the perspective it adopts. The crucial factor is whether the mathematical structure aligns with a psychological phenomenon. A useful term describing the opposite of this attitude is "*opportunist mathematics*." Stöltzner (2004) asserts that when a scientific discipline has only a weak theory of itself and poorly defined terminology, applied mathematicians adopt a strategy of *mathematical opportunism* towards this discipline. This means they engineer situations where they can apply their preferred mathematical structures to represent some phenomenon from the discipline, disregarding the phenomenon's internal structure to facilitate this engineering. Psychology, being a discipline with a weak theoretical foundation, has witnessed such engineering attempts by mathematical opportunists especially when it comes to statistics—an example is Charles Sperman who did not verify whether the attributes he considered quantitative truly possessed a quantitative structure (Michell, 2023). However, opportunism is not a characteristic exclusive to statistics. The mathematical structures presented in Lee's (2013) book may be utilized with the same degree of opportunism. In relation to psychology, opportunistic mathematics can be defined as mathematics that does not respect the structure of psychological phenomena.

Let us summarize this section: qualitative, or pure, mathematics is a discipline that seeks patterns in reproducible mental objects, sometimes employing imprecise terminology with the hope of refining it in the future. It differs from its counterpart, applied mathematics, in that it does not make compromises regarding the mutual relations of the mental objects it studies, which should be consistent with each other.

## 3 Qualitative mathematics as an alternative to measurement

Quantitative measurement attracts scholars due to its exactness, precision, rigor, and clarity (Michell, 1999:34; Gould, 1996). It also enables the standardization of processes and objective decision-making for governments (Porter, 1995). However, it has also faced sharp criticism from many scholars in psychology. Some psychologists think that quantitative models might not describe the psychological phenomena well (Guyon et al., 2018). Psychologists therefore complain that numerical measurement suitable for physics is not suitable for psychology (Trendler, 2009; Slaney, 2023), and question the application of the same rules used in physical sciences to psychology (Tafreshi, 2022). Some scholars think that regarding its mathematization, psychology should broaden its scope beyond just quantitative approaches (Omi, 2012). Michell (2003) suggests that quantitative attributes should not be the sole focus of scientific inquiry, advocating for the exploration of non-quantitative structures when evidence for quantitativeness is lacking. If no quantitative structure is

found, it should be seen as "the beginning of the search for the kind of non-quantitative structure in which nature, in this instance, is arranged" (Michell, 2003:531). Barrett (2003) suggests that graphs, language grammar or automata might be employed as non-quantitative structures when doing structural analysis of data. Some critics of measurement might view qualitative mathematics as a potential alternative, offering the same level of exactness, rigor, and clarity. I will elaborate on this possibility in the following paragraphs.

The use of qualitative mathematics, as described in Lee's (2013) book, likely requires the adoption of some form of structuralism, which posits the existence of inherent mathematical structures within the objects of psychological phenomena. In my opinion, assigning a member of a quantitative set during the (quantitative) measurement process is a similar activity to assigning a member of any other set in qualitative mathematics. The difference lies only in the type of properties that need to be evaluated. The use of qualitative mathematics would therefore require assigning elements of a structure (a set with specific properties) to certain attributes of the perceived phenomenon and evaluating whether these attributes satisfy those properties. Assigning a member of a mathematical set to some aspect of the measured phenomenon would require a human interpreter trained to conduct this measurement (Millikan, 2021). The interpreter must maintain contact with the actual phenomenon to avoid reifying the mathematical representation and using operations that are available in this representation but not applicable to the real phenomenon (Uher, 2023; Linkov, 2021).

According to metrologists, measurement needs to define the objects under measurement, the property to be measured, and the measurands. There should also be reproducibility in the measurement process—the same conditions should always produce the same measurement result. The measurement should be subject-independent, meaning the same conditions should yield the same result regardless of who is measuring (Uher, 2020). Measurement should also adhere to data generation traceability, so it should be traceable how the measurement result was produced in a specific case (Uher, 2022). In my opinion, all these requirements can be met for any mathematical structure because reproducibility, the most crucial of these requirements, is a necessary condition for something to be mathematizable. Therefore, qualitative mathematics structures might offer the same level of clarity as measurement and could serve as its alternative.

It should be noted that the term "qualitative" has different meanings in "qualitative mathematics" and "qualitative measurement" as used in metrology (Pendrill and Petersson, 2016). In metrology, "qualitative measurement" refers to simpler structures, such as nominal or ordinal scales, whereas in "qualitative mathematics," it encompasses any structure, which can be highly complex. It is also important to clarify what constitutes the similarity between "qualitative" in "qualitative research" and in "qualitative mathematics." Aspers and Corte (2019, p. 155) define qualitative research as "an iterative process in which improved understanding for the scientific community is achieved by making new significant distinctions resulting from getting closer to the phenomenon studied." In other words, it involves spending time speculating about the object being studied to uncover its specific characteristics. This is similar to mathematics, because mathematics is often considered a struggle to find the right concepts and definitions (Ziegler and Loos, 2014). The similarity between "qualitative" in "qualitative research" and in "qualitative mathematics" lies in the way researchers think, not in the structures being investigated.

While laypeople might assume that mathematics is a discipline of clear concepts and definite algorithms, a more accurate description would be a discipline that seeks to resolve ambiguities arising from incompatible frames of reference of certain concepts (which may themselves be clear). This resolution process can take hundreds of years (Byers, 2007, p. 28). The structures produced by mathematicians and the distinctions made by qualitative researchers represent two such frames of reference. Quantitative research draws much of its strength from the rigor and clarity of quantitative structures. However, if qualitative researchers hope to apply "qualitative" mathematical structures in the same straightforward manner, there is no easy solution. Establishing a correspondence between a mathematical structure and the phenomenon being studied requires a deeper understanding of both the phenomenon and the structure. Qualitative research deepens understanding of the phenomenon through the research process (Aspers and Corte, 2019), while gaining a deeper grasp of mathematical structures may require education in these structures.

## 4 Mathematical intuition might need changes in education

A crucial question concerning the use of qualitative mathematics in the social sciences is how to determine whether there is a mathematical structure that can effectively represent a social science phenomenon. This process is akin to searching for a morphism between the mathematical structure and the internal structure of the phenomenon, which would formalize the phenomenon. It is unlikely that any algorithm exists for conducting such a formalization. Insights from practicing mathematicians suggest that finding such a connection between two structures requires intuition. A scientist often spends time studying the phenomenon until inspiration strikes suddenly and unexpectedly (Hadamard, 1945; Fitzgerald and James, 2007). Creating mathematical knowledge involves "guessing a web of ideas, and then progressively strengthening and modifying the web until it is logically unassailable" (Ruelle, 2007, p. 114). To make educated guesses about the connections between mathematical and psychological structures using this intuition, a social scientist needs experience with qualitative mathematics, which is often lacking. High school students are predominantly taught quantitative mathematical disciplines, leading them to equate mathematization and formalization with quantification. Current high school curricula, such as those described by Jeřábek et al. (2021), are designed for technical fields and natural sciences, where quantification is suitable. However, the non-numerical qualitative mathematics that could be relevant for the human sciences is notably absent.

The use of qualitative mathematics in psychology might be facilitated if mathematics were taught as a search for rules valid within certain sets or as a study of relations between two sets. Examples of such subject matter could include teaching abstract algebra and conducting proofs to determine whether a set has the properties of a certain structure, such as a semigroup (a set with an associative binary operation), or examining morphisms between these sets. If a significant portion of high school curricula were composed of such mathematical content, graduates would

be less inclined to uncritically accept the quantification of real-world phenomena and would be more inclined to explore non-numerical formalizations.

The concept of teaching mathematics as an understanding of structures was promoted by the New Math movement (NMM), whose proponents believed that "math textbooks' and teachers' traditional reliance on memorization and regurgitation gave students a misleading sense of what mathematicians *do* and what mathematics was *about*" (Phillips, 2015:13). Consequently, they aimed to shift the school mathematics curriculum from learning skills and facts to acquiring conceptual understanding. The NMM, based on the ideas of the French Bourbaki group of mathematicians (Munson, 2010), initially found success in the 1960s in the US, France, and many European countries (De Bock, 2023; Gosztonyi, 2015; Prytz, 2020), but ultimately its reforms were unsuccessful. The NMM sought to provide people with a solid foundation in mathematics, enabling them to apply it in various jobs (Phillips, 2015:3). Perhaps the desire to be solid was the reason why the new math movement was unsuccessful, as parents resisted the changes, preferring that schools continue to focus on drilling students (p. 19).

NMM failed because its curriculum did not effectively train individuals in computation (Phillips, 2015:5), but psychology does not require such a drastic curriculum change as the cessation of computation drills. What psychology and social sciences might need is not necessarily solidity in the mathematical sense and teaching a deep understanding of structures, but rather instructing individuals to recognize the many possible sets that could serve as the mathematization of something.

## 5 Discussion

I have previously mentioned that what qualifies mathematics as qualitative, especially when used to represent psychological phenomena, is its alignment with those phenomena. If qualitative mathematics is ever to be utilized effectively, a primary issue must be addressed: How can we determine whether a certain mathematical structure is an appropriate representation of a phenomenon? There are significant debates about whether quantitative structures accurately represent psychological phenomena, and similar discussions could arise with other structures. A critical unresolved question is how to verify if ideas inspired by intuition are correct. Without an answer to this, the practical implementation of qualitative mathematics in psychology remains limited.

Qualitative (pure) mathematics is characterized by its attitude towards its subject matter. Therefore, applying qualitative mathematics in psychology involves searching for mathematical structures that match psychological phenomena. However, employing a specific mathematical structure in a manner that aligns with a psychological structure could be difficult, as we might lack a method to determine whether it truly fits. Another related issue is whether psychological phenomena should or even could be aligned with any mathematical structure at all. It's possible that there is no way to convincingly align some mathematical structures with psychological attributes or phenomena.

Psychological concepts are often vague, leading to questions about their existence and their ability to be thoroughly mathematized. It might be useful to remember that mathematical methods are tools for developing models, not direct representations of reality (Eronen and Romeijn, 2020), because mathematical models cannot perfectly represent reality (Bouleau, 2013). It is quite likely that for a large portion of psychological phenomena, there will be no suitable mathematical models, for other part, there will be a model applicable at a specific point in time, but the phenomenon will not be consistent and will vary with changes in time, and for another portion, there may be some mathematical models, but these could only be used as approximations of reality. Therefore, discussions on how to formalize and mathematize phenomena, and how to prepare students for flexibility in their formalizations, should be coupled with the understanding that it is acceptable to abandon formalization when a phenomenon may not possess the necessary regularity to be formalizable.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## Author contributions

VL: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# References


Aspers, P., and Corte, U. (2019). What is qualitative in qualitative research. *Qual. Sociol.* 42, 139–160. doi: 10.1007/s11133-019-9413-7

Barrett, P. (2003). Beyond psychometrics: measurement, non-quantitative structure, and applied numerics. *J. Manag. Psychol.* 18, 421–439. doi: 10.1108/02683940310484026

Bouleau, N. (2013). Can there be excessive Mathematization of the world?" in seminar on stochastic analysis, random fields and applications VII. Progress in probability, *67*, (Eds.), R. Dalang, M. Dozzi and F. Russo (Birkhäuser, Basel), 453–469

Boyer, C. B. (1949). The history of the calculus and its conceptual development: Dover Publications.

Byers, W. (2007). *How mathematicians think. Using ambiguity, contradiction, and paradox to create mathematics*. Princeton: Princeton University Press.

Byers, D. (2017). "Can you say what mathematics is?" in Humanizing mathematics and its philosophy. Essays celebrating the 90th birthday of Reuben Hersh. ed. B. Sriraman (Springer International Publishing), 45–60.

Cantor, G. (1895). Beiträge zur Begründung der transfiniten Mengenlehre. *Math. Ann.* 46, 481–512. doi: 10.1007/BF02124929

De Bock, D. (2023), Modern mathematics. An international movement?, Springer.

Devlin, K. (2012). Introduction to mathematical thinking. Palo Alto, California: Keith Devlin.

Eronen, M. I., and Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory Psychol.* 30, 786–799. doi: 10.1177/0959354320969876

Fitzgerald, M., and James, I. (2007). The mind of the mathematician: The Johns Hopkins University Press.

Franz, D. J. (2022). Psychological measurement is highly questionable but the details remain controversial: a response to Tafreshi, Michell, and Trendler. *Theory Psychol.* 32, 171–177. doi: 10.1177/09593543211062868

Gosztonyi, K. (2015). Tradition and reform in mathematics education during the "new math" period: a comparative study of the case of Hungary and France. Dissertation Thesis. Paris: Universite Paris Diderot.

Gould, S. J. (1996). The Mismeasure of man. W. W. Norton & Company.

Guyon, H., Kop, J.-L., Juhel, J., and Falissard, B. (2018). Measurement, ontology, and epistemology: psychology needs pragmatism-realism. *Theory Psychol.* 28, 149–171. doi: 10.1177/0959354318761606

Hadamard, J. (1945). "The Mathematician's mind" in The psychology of invention in the mathematical field (Princeton: Princeton University Press).

Hersh, R. (2014). Experiencing mathematics. What we do when we do mathematics? Providence, Rhode Island, USA: American Mathematical Society.

Higham, N. J. (2015). "What is applied mathematics?" in The Princeton companion to applied mathematics. eds. N. J. Higham, M. R. Dennis, P. Glendinning, P. A. Martin, F. Santosa and J. Tanner (Princeton University Press), 1–8.

James, W. (1950). The principles of psychology: Dover Publications.

Jeřábek, J., Krčková, S., Hučínová, L., Balada, J., Baladová, G., Boněk, J., et al. (2021). Rámcový vzdělávací program pro gymnázia [Framework curriculum for grammar schools]. Prague: Ministry of Education, Youth, and Sports of the Czech Republic.

Lee, R. (ed.). (2013). Qualitative mathematics for the social sciences. Mathematical Models for Research on Cultural Dynamics. Routledge.

Linkov, V. (2021). Research based on scientific realism should not make preliminary assumptions about mathematical structure representing human behavior: Cronbach and Gleser's measure as an example. *Theory Psychol.* 31, 465–470. doi: 10.1177/0959354321101608

Michell, J. (1999). Measurement in psychology. Critical History of a Methodological Concept. Cambridge University Press.

Michell, J. (2003). Epistemology of measurement: the relevance of its history for quantification in the social sciences. *Soc. Sci. Inf.* 42, 515–534. doi: 10.1177/0539018403424004

Michell, J. (2023). Professor spearman has drawn over-hasty conclusions: unravelling psychometrics Copernican revolution. *Theory Psychol.* 33, 661–680. doi: 10.1177/09593543231179446

Millikan, R. G. (2021). Neuroscience and teleosemantics. *Synthese* 199, 2457–2465. doi: 10.1007/s11229-020-02893-9

Munson, A. (2010). Bourbaki at seventy-five: its influence in France and beyond. *J. Math. Educ. Teachers College* 2010, 18–21. doi: 10.7916/jmetc.v1i2.686

Omi, Y. (2012). Tension between the theoretical thinking and the empirical method: is it an inevitable fate for psychology? *Integr. Psych. Behav.* 46, 118–127. doi: 10.1007/s12124-011-9185-4

Pendrill, L., and Petersson, N. (2016). Metrology of human-based and other qualitative measurements. *Meas. Sci. Technol.* 27:094003. doi: 10.1088/0957-0233/27/9/094003

Phillips, C. J. (2015). The new math. A political history: The University of Chicago Press.

Porter, T. M. (1995). Trust in Numbers. The Pursuit of Objectivity in Science and Public Life. Princeton University Press.

Prytz, J. (2020). The OECD as a booster of National School Governance. The case of new math in Sweden, 1950-1975. *Foro de Educación* 18, 109–126. doi: 10.14516/fde.824

Ruelle, D. (2007). The Mathematician's brain: Princeton University Press.

Schwartz, J. (2006). The pernicious influence of mathematics in science, 18 unconventional essays on the nature of mathematics, (Ed.) R. Hersh (Springer), 231–235

Slaney, K. L. (2023). Why force a square peg into a round hole? The ongoing (pseudo-) problem of psychological measurement. *Theory Psychol.* 33, 138–144. doi: 10.1177/09593543221128522

Stöltzner, M. (2004). On optimism and opportunism in applied mathematics: mark Wilson meets John von Neumann on mathematical ontology. *Erkenntnis* 60, 121–145. doi: 10.1023/B:ERKE.0000005144.79761.02

Tafreshi, D. (2022). Sense and nonsense in psychological measurement: a case of problem and method passing one another by. *Theory Psychol.* 32, 158–163. doi: 10.1177/09593543211049371

Toomela, A. (2008). Variables in psychology: a critique of quantitative psychology. *Integr. Psych. Behav.* 42, 245–265. doi: 10.1007/s12124-008-9059-6

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory Psychol.* 19, 579–599. doi: 10.1177/0959354309341926

Uher, J. (2020). Measurement in metrology, psychology and social sciences: data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Qual. Quant.* 54, 975–1004. doi: 10.1007/s11135-020-00970-2

Uher, J. (2021). Psychometrics is not measurement: unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.* 41, 58–84. doi: 10.1037/teo0000176

Uher, J. (2022). Functions of units, scales and quantitative data: fundamental differences in numerical traceability between sciences. *Qual. Quant.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Soc. Personal. Psychol. Compass* 17:e12740. doi: 10.1111/spc3.12740

Ziegler, G.M., and Loos, A. (2014). Teaching and learning "What is Mathematics", Proceedings of the International Congress of Mathematicians Seoul 2014. IV. Invited Lectures, (Eds) S. Y. Jang, Y.R. Kim, D.W. Lee and I. Yie, Kyung Moon SA, 1203–1216

frontiers | Frontiers in Psychology

TYPE Methods
PUBLISHED 04 October 2024
DOI 10.3389/fpsyg.2024.1419564



OPEN ACCESS

EDITED BY
Jan Ketil Arnulf,
BI Norwegian Business School, Norway

REVIEWED BY
Kjell Ivar Øvergård,
University of South-Eastern Norway, Norway
Prosper Kwei-Narh,
Inland Norway University of Applied Sciences, Norway

*CORRESPONDENCE
Philipp Brauner
philipp.brauner@rwth-aachen.de

RECEIVED 11 June 2024
ACCEPTED 04 September 2024
PUBLISHED 04 October 2024

CITATION
Brauner P (2024) Mapping acceptance: micro scenarios as a dual-perspective approach for assessing public opinion and individual differences in technology perception. *Front. Psychol.* 15:1419564. doi: 10.3389/fpsyg.2024.1419564



# Mapping acceptance: micro scenarios as a dual-perspective approach for assessing public opinion and individual differences in technology perception

Philipp Brauner*

Communication Science, RWTH Aachen University, Aachen, Germany

Understanding public perception of technology is crucial to aligning research, development, and governance of technology. This article introduces micro scenarios as an integrative method to evaluate mental models and social acceptance across numerous technologies and concepts using a few single-item scales within a single comprehensive survey. This approach contrasts with traditional methods that focus on detailed assessments of as few as one scenario. The data can be interpreted in two ways: Perspective (1): Average evaluations of each participant can be seen as individual differences, providing reflexive measurements across technologies or topics. This helps in understanding how perceptions of technology relate to other personality factors. Perspective (2): Average evaluations of each technology or topic can be interpreted as technology attributions. This makes it possible to position technologies on visuo-spatial maps to simplify identification of critical issues, conduct comparative rankings based on selected criteria, and to analyze the interplay between different attributions. This dual approach enables the modeling of acceptance-relevant factors that shape public opinion. It offers a framework for researchers, technology developers, and policymakers to identify pivotal factors for acceptance at both the individual and technology levels. I illustrate this methodology with examples from my research, provide practical guidelines, and include R code to enable others to conduct similar studies. This paper aims to bridge the gap between technological advancement and societal perception, offering a tool for more informed decision-making in technology development and policy-making.

KEYWORDS

**cognitive maps, technology acceptance, public perception, micro scenarios, psychometric paradigm, mental models, attributions, survey methodology**

## 1 Introduction

Technological advancements are often accompanied by dilemmas and they must be aligned with human norms and values. History has many instances of such ethical dilemmas, such as mechanization and industrialization, leading to enhanced productivity but also accompanied by substandard working conditions (Engels, 1845; Watt, 1769), movable types and the printing press yielding increased literacy but resulting in the dissemination of pamphlets containing misinformation (Steinberg, 1974; Eisenstein, 1980), and the invention of clothing for protection and warmth leading to the environmental repercussions of fast fashion, causing ecological damage (Kvavadze et al., 2009; Niinimäki et al., 2020).

When technologies become a part of our life, it is essential to integrate the perspective of us—the people—to understand how we evaluate them, what we attribute to them, and how they relate to our norms and values (Guston and Sarewitz, 2002; Rogers et al., 2019; Lucke, 1995). When technologies reflect peoples' values, they are more likely to be accepted, adopted, and integrated into daily life. Conversely, if a technology conflicts with prevailing values, it may face resistance or rejection. However, technology may change our norms and values and our norms and values may shape how a technology is used. For instance, the Internet has fostered values of openness and connectivity, while these values have, in turn, driven the development of social media platforms. Similarly, technologies can afford new possibilities that lead to the development of new values. For example, the rise of renewable energy technologies has spurred values around environmental sustainability. However, there are instances where technologies and values are in opposition. Surveillance technologies, for example, clash with values of privacy and individual freedom. Also, technologies often introduce ethical dilemmas where existing values are challenged, such as the advent of genetic editing technologies like clustered regularly interspaced short palindromic repeats (CRISPR) raises questions about the value of human life and natural processes.

There are various methods for assessing peoples' perception of technologies: ranging from scenario-based approaches, over living labs, to hands-on experiences with readily available technologies (Tran and Daim, 2008; Grunwald, 2009). The majority of empirical approaches use different concepts of technology acceptance to assess specific technologies and systems. Referring to model-based approaches, the constructs behavioral intention to use and actual use are often applied to measure technology acceptance (Davis, 1989; Marangunić and Granić, 2015). Other approaches focus more on affective evaluations, addressing the social perception of specific technologies and systems (Agogo and Hess, 2018; Zhang et al., 2006). Furthermore, the evaluation of single technologies often contains a modeling and trade-off between specific technology-related perceived (dis-)advantages affecting the final evaluation and acceptance (Buse et al., 2011; Offermann-van Heek and Ziefle, 2019).

Although research on technology acceptance and evaluation has increased significantly in the last decades, the majority of the studies focus on the evaluation of single applications (Rahimi et al., 2018; Al-Emran et al., 2018) describing specific requirements, benefits, and barriers of its usage in depth. In contrast, a broader view on diverse technologies' assessment enabling a comparison and meta-perspective on a variety of technologies has rarely been realized so far. Further, most evaluations based on conventional acceptance models or their adaptations do not facilitate mapping or contextual visualization of a wider range of technologies and concepts.

Therefore, this article aims at presenting a novel micro-scenario approach, enabling a quantitative comparison of a broad variety of technologies, applications, or concepts based on affective evaluations, in parallel with the interpretation of an individual's assessment as individual dispositions, as well as a concept of visualizing the evaluations as visual cognitive maps.

The article is structured as follows: Section 1 provides the introduction and motivates this methodological approach. Section 2 reviews the current state of technology acceptance evaluations and related measures, highlighting existing research gaps. Section 3 defines micro-scenarios as an integrated contextual perspective and discusses the strengths and limitations of this approach. Section 4 introduces guidelines and requirements for designing surveys based on micro-scenarios. Section 5 presents a concrete application example, showcasing the results of a recent study on the acceptance of medical technology. This example demonstrates the practical value of the approach and the insights it can provide (all data and analysis code are available as open data). Section 6 concludes with a summary and a discussion of the methodological strengths and limitations of the approach, as well as its overall usefulness. Finally, the Appendix details the technical implementation of micro-scenario-based surveys, along with actionable examples and R code for conducting similar studies.

## 2 Background and related measures

The following section presents the theoretical background and introduces related empirical concepts and approaches, as well as related methodological procedures.

### 2.1 Related concepts and approaches

A fundamental concept in acceptance research is mental models. These are simplified, cognitive representations of real-world objects, processes, or structures that enable humans and other animals to evaluate the consequences of their (planned) actions. These simplified models influence our behavior (Jones et al., 2011; Johnson-Laird, 2010; Craik, 1943): When aligned with reality, they facilitate efficient and effective interactions with the surroundings (Gigerenzer and Brighton, 2009). Conversely, erroneous mental models restrict the correct assessment of the environment and hinder accurate inferences (Gilovich et al., 2002; Breakwell, 2001).

Extracting mental models through empirical research provides insights into how basic attitudes and attributions are shaped and change.

For this purpose, many qualitative (for example, interviews and focus groups or rich picture analysis) and quantitative approaches (for example, surveys or experimental studies) are available. One frequently used method in acceptance research involves scenarios depicting technologies or their applications, which are integrated in qualitative, quantitative, or mixed-methods approaches (Kosow and Gaßner, 2008). In this approach, a new technology or service is described textually and/or visually within a scenario and then evaluated by study participants based on various criteria. Typically, these scenarios are designed to let participants evaluate a single technology, application, or situation in detail. Only occasionally, a few (rarely more than three) different technologies or their applications are assessed. Through these scenarios, participants evaluate their perceptions, attitudes, and acceptance of the specific research object. While these responses are not the mental models, they reflect the participants mental models.

There are multiple ways to describe the perception of technologies and the influencing factors involved. A prominent example are studies based on the technology acceptance model

(TAM) or the increasingly specific models derived from it (Davis, 1989; Rahimi et al., 2018). TAM postulates that the later actual use of a technology—originally office applications—can be predicted in advance via a model of the individuals' perceived ease of using the system, and the perceived usefulness, and the intention to use the system. Later models have extended the concept of predicting the later use through the usage intention and an increasingly diverse set of antecedents. Examples include the hedonic value of a product, or if others could provide support in case of troubles (Venkatesh et al., 2012). Nowadays, new models are being proposed for each seemingly new technology; but rarely are different technologies compared in a single study. While the core idea remains the same—predicting use by linking intention to use to other factors—there are now many an overwhelming number of models and constructs used in technology acceptance research (Marikyan et al., 2023 gives a meta-review on the constructs used in 693 studies).

As not every technology is *used* by individuals (such as a nuclear power plant), other models focus on other outcome variables. For example, the value-based acceptance model shares many similarities with the TAM (Kim et al., 2007), yet it focusses on a *perceived value* of the evaluated entity instead of the *intention to use* (and *use*). Again, different predictors are related to the valence as the target variable and researchers can weight the factors that influence to higher or lower valence of a topic.

A common feature of all these approaches is that one or very few technologies or scenarios are assessed in detail. In contrast, the micro-scenario approach looks at many different scenarios and tries to put them in relation to each other and to uncover connections and differences between the scenarios.

Beyond the need to better understand technology attributions and acceptance at both technological and individual levels, there is also a need to enhance our methodological tools. Studies suggest that questionnaires assessing technology acceptance (and likely other questionnaires) may be biased due to the lexical similarity of items and constructs (Gefen and Larsen, 2017). A significant portion of the TAM can be explained solely through linguistic analysis and word co-occurrences (although subjective evaluations further improve the model). To further develop and validate our methods, it is essential to consider different and new perspectives on the phenomena we study (Revelle and Garner, 2024).

## 2.2 Related methods

This section presents existing and partly related methodological procedures in order to identify similarities, but also differences and gaps, the approach presented here addresses.

### 2.2.1 Vignette studies

At first sight, vignette studies are related to this approach, although they are rather the opposite of the method presented here. Vignette studies are a way to find out which characteristics influence the evaluation of people, things, or services. Essentially, in vignette studies, a base scenario is parameterized using certain dimensions of interest, displayed and evaluated by subjects based on one or more evaluation dimensions. Examples include studies on the influence of cognitive biases in evaluating job applications: The same job applications may be framed by the applicants' age, ethnicity, or social group and as target variable, for example, the likelihood of interviewing the person for a job is measured (Bertogg et al., 2020). This approach enables to examine which factors have an influence on, for example, the likelihood to get invited to the job interview and also to quantify the weight of each factor using, for example, linear regressions on the factor that constitute the vignettes (Kübler et al., 2018). The key difference between the established vignette studies and the approach presented here is that vignette studies aim at identifying influencing factors for one particular entity (e.g., an applicant) while micro-scenarios address the influencing factors of different topics in one shared research space.

### 2.2.2 Conjoint analysis

There is also a similarity to the conjoint analysis (CA) approach. CA were developed in the 1960s by Luce and Tukey (1964) and are most prevalent in marketing research. Participants are presented a set of different products that are composed of several attributes with different levels. Depending on the exact methods, they either select the preferred product out of multiple product configurations, or decide whether they have a purchase intention for one presented option. CA results in a weighting of the relevant attributes for production composition (e.g., that car brand may be more important than performance or color) and the prioritizations of the levels of each attribute (e.g., that red cars are preferred over blue ones). While this approach shares some similarities with the micro-scenarios (e.g., systematic configuration of the products resp. scenarios) there are also differences. A key difference is that CA has one target variable (e.g., selection of the preferred product), whereas the micro-scenarios have multiple target variables and each scenario is evaluated. Furthermore, CA has tools for calculating optimal product configurations and market simulators. While the market simulation allows a comparison of multiple actual or fictitious products, it does not facilitate the identification of blank areas in a product lineup or how the products relate to each other beyond a unidimensional preference. Also, while results from a CA can be used to define customer segments by means of a latent class analysis, the individual preferences can not easily be interpreted as personality factors.

### 2.2.3 (Product) positioning

Another similar approach is "positioning" in marketing (Ries and Trout, 2001), in which products and brands in a segment are evaluated in terms of various dimensions and presented graphically. Based on the graph, new products or brands can be developed to fill gaps or reframed and thus moved to different positions. However, the approach presented here does evaluate and map topics. It focuses on an understanding of the public perception of topics, it does not aim to create new topics, and the evaluated topics can usually not easily be changed (i.e., power plant technologies). Furthermore, beyond the positioning, it does not aim at modeling or explaining the role of individual differences in the evaluations.

### 2.2.4 Psychometric paradigm of risk perception

There are similarities between the micro-scenario approach and Slovic's psychometric paradigm and his seminal works on risk perception (Slovic, 1987; Fischhoff, 2015). Based on the analysis of individual studies, his work suggests that risk attributions have a two-factor structure, with dread risk and unknown risk identified by factor analysis. He used these two-dimensional factors to map a variety of different hazards on a scatterplot ("cognitive map") that looks very similar to the visual outcomes of the micro-scenario approach. However, Slovic's approach focusses more on the psychological aspects of how people perceive and categorize risks and it's based on many individual studies. In contrast, micro-scenarios are more pragmatic and allow arbitrary evaluation dimensions. Building on a single integrated survey and considering risk, utility, or other relevant dimensions can inform researchers, decision-makers, and policy makers in a tangible and applicable manner.

### 2.2.5 Experimental factorial designs

A common theme in psychological research is factorial designs that involves manipulating two or more independent variables simultaneously to study their combined effects on one or more dependent variables (Montgomery, 2019; Field, 2009). It allows us to examine and weight the influence of the factors and the interaction effects between multiple factors (Montgomery, 2019). This concept is extensively and predominantly used in experimental cognitive and behavioral research. However, its application in scenario-based acceptance studies is limited. When used in such studies, they typically only involve single factors due to the large number of dependent variables queried, which would otherwise make the surveys unmanageable.

## 2.3 Similarities and methodological gap

Summarizing these different methodological approaches, they indeed share similarities with the approach presented here. However, they differ in terms of the usage context and purpose of use, variable reference and scope, their target size and their comparability. What is still needed is a broader view of the assessment of on diverse technologies, enabling a comparison and meta-perspective on a variety of technologies enabling comparative mappings or visualizations.

Therefore, a novel micro-scenario approach is introduced in the following section. In the single survey, this approach allows both the assessment and comparison of a wide range of topics, applications, or technologies, as well as the measurement of individual differences in the assessments based on affective evaluations.

# 3 Micro-scenarios as an integrated contextual perspective

The goal of the micro-scenario approach is to gather the evaluation of a *wide range of topics or technologies* on *few selected response variables* and put the different evaluations into context. Hereto, the subjects are presented a large number of different short scenarios and how they evaluate those scenarios is measured using a small set of response variables. The scenario presentation can be a short descriptive text, and/or images, or, in extreme cases, just a single word about an evaluated technology or concept. The former offers the possibility to give some explanation on each of the evaluated topics, whereas the latter essentially measures the participants' affective associations toward a single term. Section 4.1 outlines guidelines for creating the set of scenarios.

Each scenario is then evaluated on the same small set of response items. Which dimensions are used for the assessment depends on the specific research question and may, for example, be risk, benefit, and overall evaluation of a technology to identify (in-)balances in risk-utility tradeoffs (cf. Fischhoff, 2015), the intention to use and actual use of technology as in the TAM (cf. Davis, 1989) to identify different motives for using software applications, the perceived sensitivity of data types and the willingness to disclose the data to others to understand the acceptance barriers to personal life-logging and monitoring at the work-place (cf. Tolsdorf et al., 2022), or other dependent variables that match the research focus. I suggest the use of only single item-scales and only to measure the most relevant target dimensions (Fuchs and Diamantopoulos, 2009; Ang and Eisend, 2017; Rammstedt and Beierlein, 2014). Typically, one would use three to five items for the evaluation of each micro-scenario. On the one hand, this sacrifices the benefits of psychometric scales with high internal reliability. On the other hand, this offers the benefits that (a) each scenario can be evaluated quickly and cost-effective (Woods and Hampson, 2005; Rammstedt and Beierlein, 2014) , (b) perceived repetitiveness of psychometric scales is avoided and the survey can be more interesting for the participants, and (c) many scenarios can be evaluated in a single survey. Section 4.2 details the selection of suitable items. Figure 1 illustrates this concept.

With a suitable combination of scenarios and dependent variables, the approach offers two complementary research perspectives:

*Perspective 1:* As the first research perspective, the evaluations can be understood as user variables (individual differences between the participants) and correlations between age, gender, or other user factors can be investigated. The evaluation of various topics can essentially be considered as a repeated reflexive measurement of the same underlying latent construct (see Figure 2).

*Perspective 2:* As the second research perspective, the evaluations serve as technology evaluations and relationships between the evaluation dimensions across the different topics can be studied (differences and communalities between the queried topics) (see Figure 3).

This approach has three distinct advantages:

*Efficient evaluations:* One advantage lies in a pragmatic and efficient evaluation of the topics by the participants, as the cognitive effort required to evaluate the topics is comparably low. Following the mainstream model or answering survey items participants (1) need to understand the question, (2) gather relevant information from long-term memory, (3) incorporate that information into an assessment, and (4) report the resulting judgment (Tourangeau et al., 2000). Here, the respondents have to retrieve their attitude toward each topic only once and then evaluate it on a repeating set of the same response items that should be presented in the

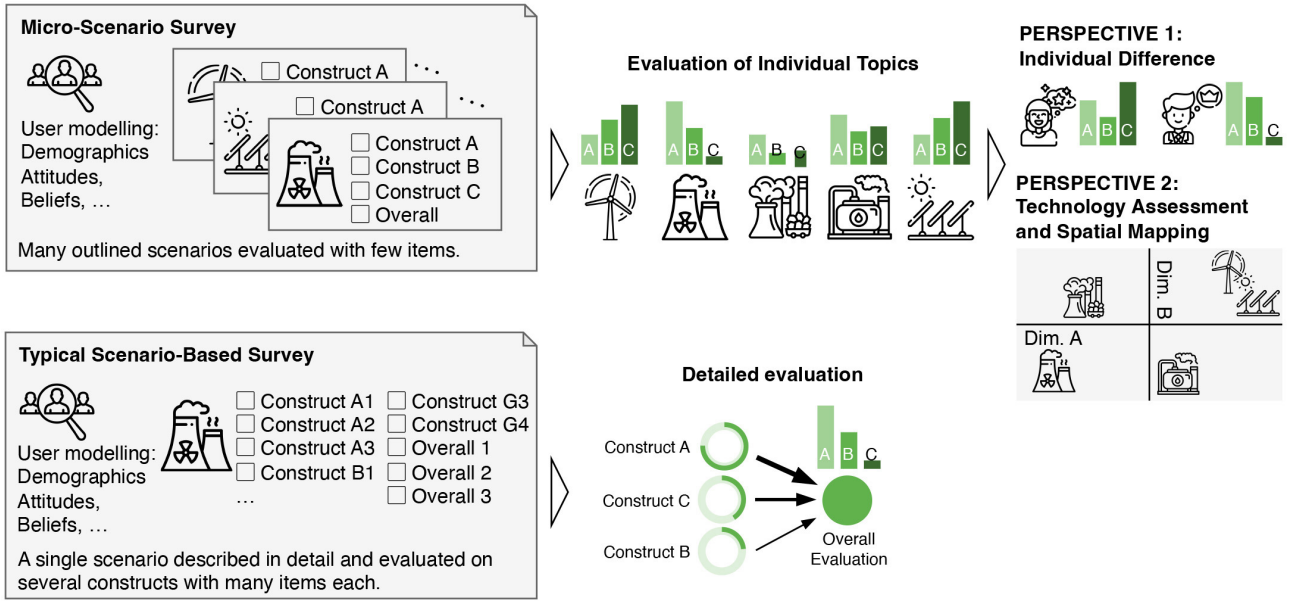



FIGURE 1
Concept of a micro-scenario survey compared to a conventional scenario-based survey.

same order. While the number of items in these studies is high, its repetitive structure responds to them cognitively easily. That facilitates assessing large number of topics within the same survey.

*Joint evaluations:* In addition, a large number of different topics can be analyzed in a single integrated study. Based on the selected dependent variables for the topics, the relationships among these can further be studied. In a study, we used a linear regression analysis to study the influence of perceived risk and perceived utility on the overall valence of medical technology (Brauner and Offermann, 2024, see the example in Section 5). Based on the calculated regression coefficients and with a high explained variance (≫90%), we could argue that the variance in overall evaluation of medical technology is mostly determined by the perceived benefits rather than the perceived risks.

*Visual interpretation:* Furthermore, the multivariate scenario evaluations can be put into context and presented on two-dimensional spatial maps enabling a visual interpretation of the findings (see Figure 4 for an abstract example and Figure 3 for a view on the required data structure). This representation facilitates the analysis of the spatial relationships between the topics and the identification of topics that diverge from others and thus require particular attention by the public, researchers, or policymakers. To stay in the aforementioned example, we mapped the risk-utility tradeoff across a variety of different topics (see Figure 6 in Section 5). This *visual mapping* of the outcomes can then be interpreted as follows: *First*, one can interpret the breadths and position of the distribution of the topics on the *x*- or *y*-axis. A broader distribution suggests a more diverse evaluation of the topics, whereas a narrow distribution is an indicator for a rather homogenous evaluation. The mean of the distribution of the topics indicates if the topics are—on average—perceived as useful or useless. *Second*, the slope and the intercept of the resulting regression line can be interpreted: The steepness of the slope indicates the tradeoffs between the two mapped variables. If the slope is +1, an increase by one unit of perceived utility means an increase by one unit of perceived risk. Steeper or flatter slopes indicate different tradeoffs. *Third*, one can inspect the position of the individual topics on the map. Elements from left to right are perceived as having less or more risk. Elements from the top to bottom are perceived as having higher to lower utility. Consequently, elements on or near the diagonal are topics where risk and utility are in balance. While some topics are perceived as more and others as less risky, the additional perceived risk is compensated by additional utility. However, if elements are far off the diagonal, there is perceived risk and the utility is unbalanced, potentially because a minor utility does not compensate for a higher degree of risk. Hence, these topics require particular attention from individuals, researchers, or policymakers.

Obviously, other research questions may build on different pairs of dependent variables to be mapped, such as intention and behavior, the same dependent variable by different groups, such as experts and laypeople, or usage contexts, such as passive and active use of technology.

In summary, the micro-scenario approach captures the individual participants' attributions toward various topics but instead of considering these only as individual differences, they are also interpreted as technology attributions and analyzed accordingly.

Consequently, I define micro-scenarios as a methodological approach that facilitates the comprehensive assessment of numerous technologies or concepts on few response items within a single survey instrument. This method enables the quantitative analysis and visual illustration of the interrelationships among the technologies or concepts being investigated. Furthermore, micro-scenarios enable the interpretation of the respondents' overall attributions as personal dispositions, thereby providing insight into individual perceptions and beliefs.

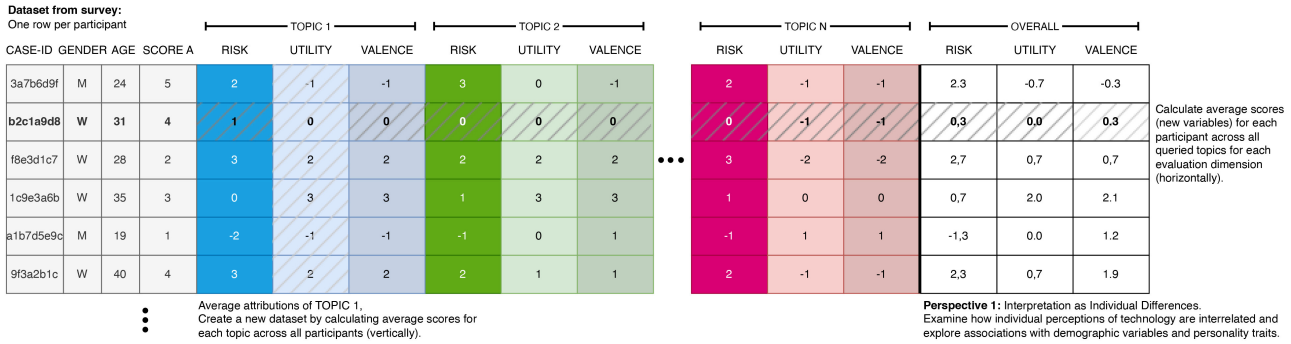



FIGURE 2
Illustration of the two different sets of aggregatable variables: first, the attributions to the queried topics along the evaluation dimensions (vertically across the dataset) defining the attributions toward the topics. Second, the average ratings of the topics per participant (horizontally in the dataset), defining a personality state or trait.

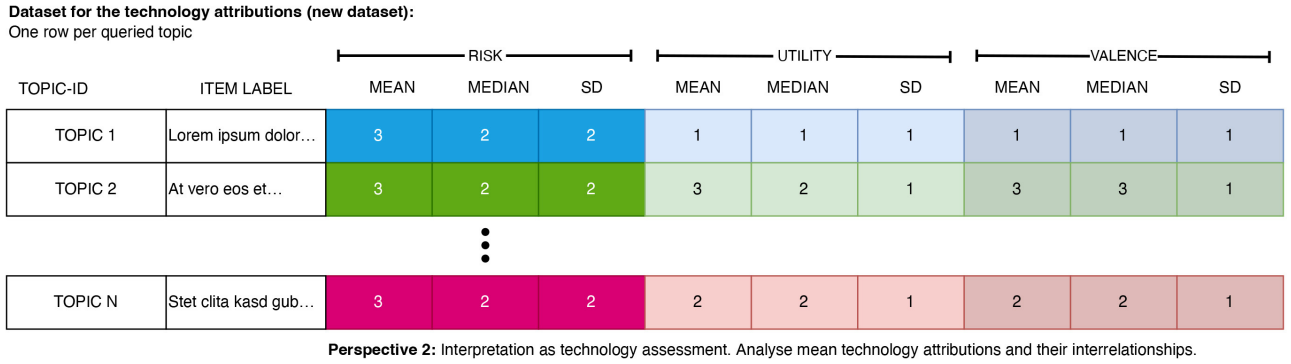



FIGURE 3
IResulting data format for the evaluation of the topics. Each row stores the mean evaluation (and its dispersion) of a topic. This data can be further analyzed using correlation or regression analyses.

# 4 How to conduct micro-scenario surveys

This section outlines the guidelines for conceptualizing a micro-scenario study. Hereby, three areas have to be considered. First, the identification and definition of a suitable research space. Second, the definition of suitable dependent variables that are relevant, suitable for visual mapping, and facilitate further analyses. Finally, the identification of additional variables for modeling the participants that can then be related with the aggregated topic evaluations. In the following, I discuss each point briefly and provide a few suggestions. Obviously, this can neither replace a text book on empirical research methods (e.g., Döring, 2023; Groves et al., 2009; Häder, 2022) nor a systematic literature review on current research topics. However, it should give some guidance on which aspects need to be considered to create an effective survey.

## 4.1 Defining the scenario space

Researchers first have to define the general research domain (such as the perception of Artificial Intelligence, medical technology, or energy sources). Technologies that serve similar functions or are used in similar contexts can be compared in terms of public perception and value alignment. For example, different renewable energy technologies can be compared based on values related to environmental impact and sustainability. However, technologies serving fundamentally different purposes may be less comparable and thus the micro-scenario approach is then not suitable: Comparing an entertainment technology like virtual reality to a healthcare technology like MRI machines may not yield meaningful insights due to the divergent values and expectations involved. Based on this, a set of suitable scenarios needs to be identified. To compile the set of scenarios there are two different approaches:

On the one hand, the scenarios can be defined intuitively, based on the results of an extensive literature review, or as a result of appropriate preliminary studies [such as interviews or focus-groups (Courage and Baxter, 2005)]. However, this bears the risk that the selection of queried topics is neither random nor systematically constructed. While the analysis can yield interesting results, there is a risk that the findings may be affected by a systematic bias (for example, Berkson, 1946's paradox, where a biased sample leads to spurious correlations).

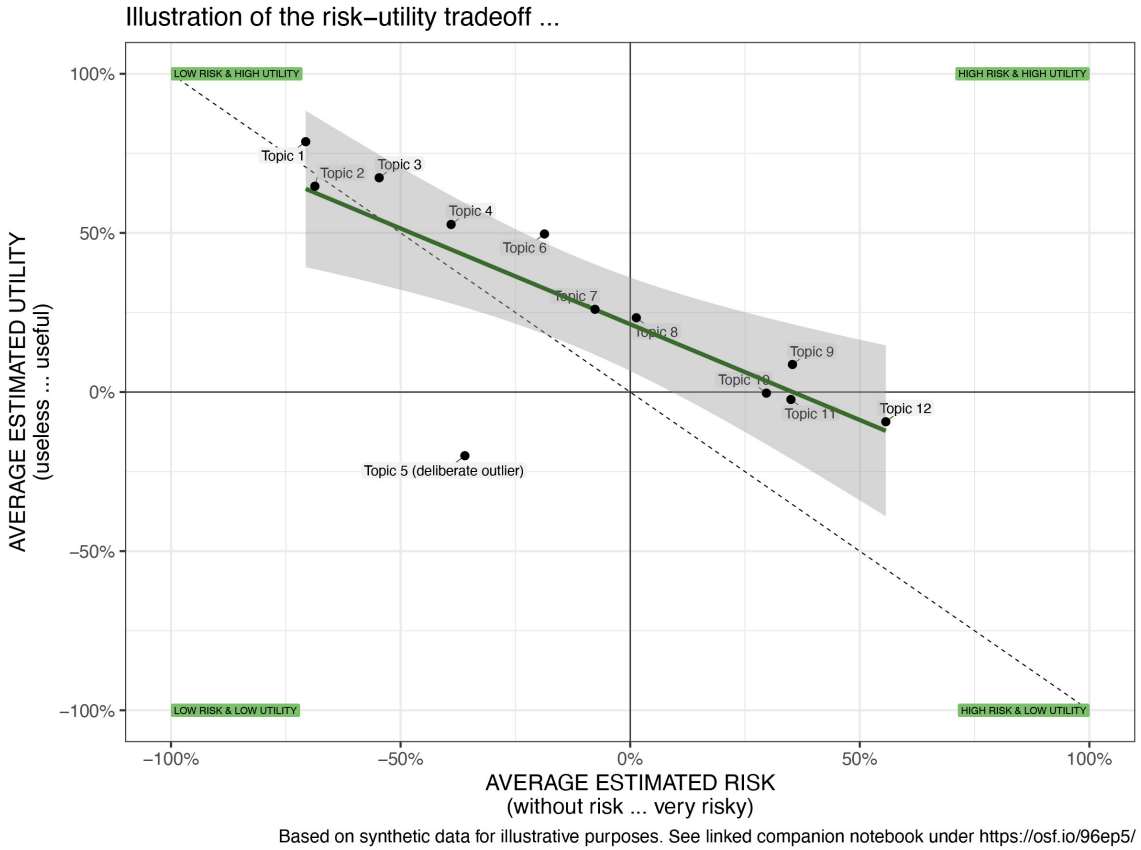



FIGURE 4
Illustrative example of the risk-utility tradeoff of technologies. Apparently, both evaluation dimensions are inversely correlated (negative slope). Some topics are perceived as risky while others are perceived as safe. Both the utility and risk distributions are above the a neutral judgement, meaning that most topics are perceived as risky and useful.

On the other hand, systematic biases can be avoided if the topics for the research space are compiled systematically. If possible, I recommend identifying an underlying factorial structure of the research space and exploring the research space systematically by querying 1 to *N* topics for each linear combination of this space (i.e., latin square design). For example, if one wants to evaluate different forms of energy generation, one could first identify possible factors of an underlying design space of the topics (e.g., size, sustainability, risk, co-location with housing, ...) and their respective levels (e.g., ranging from small to large, not sustainable to circular, ...). Next, and based on the latin square method, suitable instances for each of the factor combinations can be identified and selected. This avoids that some areas of the underlying research space are over and others are under represented in the sample of topics. Hence, this approach reduces systematic bias in the data due to non-biased sampling of the topics.

Based on conducted studies, I suggest querying about 16–24 topics, to balance the expressiveness of the results with the length of the survey and to avoid the effects of both learning and fatigue. If more topics need to be queried, one can use *randomized sampling* of the queried topics: While many topics are in the survey, only a random subset is rated by each participant. Note however, that random sampling of technologies or topics may have unindented side-effects that may limit the validity of the study due to the risk of biased sampling.[1]

What suitable dimensions for the research space are, depends on the general research domain. As outlined above, a research space for energy conversation technology may build, for example, on the dimensions of degree of sustainability, price, size, or decentralization. A research space for medical technology (see Section 5) may build on the dimensions how invasive a technology is, how digital it is, whether it is used by patients or doctors. Beyond that one can also include other factors, such as when a topic or technology became public (cf. Protzko and Schooler, 2023).

Beyond the underlying factorial structure, the selected scenarios should otherwise be comparable. Participants should evaluate different instances of a technology and not hard to compare concepts. Of course, the scenario descriptions should be developed and iteratively refined to ensure comprehensibility for the participants and to facilitate the evocation of a mental model among the participants.

1 To mitigate this, one should build on a sufficiently large subset of technologies and a larger sample of participants. Further, one might consider suitable data imputation strategies.

## 4.2 Defining the topic evaluation variables

Next, the appropriate dependent variables for the assessment of the topics need to be identified. Of course, this depends on the selected research context and the targeted participants of the survey. For example, medical and biotechnologies often involve ethical considerations and personal values related to life, health, and body autonomy. Information and communication technologies (ICTs) influence and are influenced by values related to privacy, freedom of expression, and information accessibility. Hence, this article only provides some more general remarks on this selection: First, the article exemplifies the selection of variables by sketching three potential research questions. Secondly, it discusses how many and which items can be used for operationalization. Finally, it suggests how the reliability of the measurement can be checked.

For example, to study risk–benefit trade-offs and their relation to the willingness to accept or adopt a technology (Fischhoff, 2015; Brauner and Offermann, 2024), one might to query the *perceived risk*, the *perceived benefit*, and the overall acceptance or *willingness to adopt* a technology (Davis, 1989). This would allow to calculate a multiple linear regression (across the average topic evaluations) with the average risks and benefits of the technologies as independent variables and the willingness to adopt as dependent variable. For technologies that are not adopted by individuals (e.g., different types of power plants), an overall *valence* might be more suitable (Kim et al., 2007). In a different study, one might be interested in the *perceived sensitivity* and the *willingness to disclose* the information from various sensor types for personal life-logging (Lidynia et al., 2017) or workplace monitoring (Tolsdorf et al., 2022).

Suitable dependent variables can be adapted from other research models. For example, to evaluate a number of different mobile applications, one might refer to technology acceptance model (see above) and its key dimensions *perceived ease of use*, *perceived usefulness*, and *intention to use* or actual *use*. If the perception of risk and *benefits* (or utility) is of interest, one may consider *risk* and *benefits* as target variables: In one study, colleagues and I build on Fishhoff's psychometric model of risk perception (Fischhoff, 2015) to study risk-benefit tradeoffs in the context of Artificial Intelligence (AI) (Brauner et al., 2024): For a large number of developments and potential implications of AI, we wanted to explore if the overall evaluation is rather driven by the perceived risks or by the perceived benefits. Hence, we measured the overall evaluation as *valence* (positive—negative), the perceived *risk* (no risk—high risk), and the perceived *utility* (useless—useful).[2] Furthermore, one might study the *intention-behavior*-gap in different contexts.

In a recent study, an attempt was made to measure perceived expectancy, which refers to whether participants believed a presented development is likely to occur in the future (Brauner et al., 2023). However, no relationship to other variables in the study could be identified. This corroborates that forecasting seems to be difficult, especially for laypeople (Recchia et al., 2021).

The number of queried items for each topic should be limited. As the number of dependent variables for each topic increases the survey duration linearly, this can quickly lead to excessively long questionnaires. Hence, I am proposing to use single item scales for each relevant target dimension (Woods and Hampson, 2005; Rammstedt and Beierlein, 2014; Fuchs and Diamantopoulos, 2009). Consequently, I advise building on the existing research models and select the items that were identified as working particularly well in other studies (e.g., select the item with the highest item-total-correlation (ITC) from well-working scales).

In previous studies building on this approach, the number of target dimensions varied between two and seven. A number between three to five was working particularly well and should work for many contexts (e.g., to study the relationship between risk, benefit and acceptance, or intention and behavior).

Using a semantic differential for querying the dependent variables is suggested. These have metric properties and usually require low cognitive effort by the participants, as these items can usually be more easily interpreted, evaluated, and the appropriated response be selected (Messick, 1957; Woods and Hampson, 2005; Verhagen et al., 2018). Especially as the participants report on a larger number of scenarios and items, I suggest to keep the items and the response format as easy as possible.

## 4.3 Modeling the influence of user diversity

Finally, one should consider the choice of additional user variables that should be surveyed and related to the topic evaluations. Beyond the usual demographic variables, such as age and gender, this strongly depends on the specific research questions and context of the study. Hence, I can only provide some general ideas and remarks.

The first perspective of this approach facilitates the calculation of mean topic evaluations, for example, the mean valence or the mean risk attributed to the topics (see Figure 5). These calculated variables can then be considered as personality states (changing over time) or traits (relatively stable), and can be related to the additional user variables.

Hence, one should assume relationships between the newly calculated variables from the topic evaluation and the additional user variables. In the case of the study on the perception of AI, the average assessments across the topics (see Section 4.2) *valence*, *risk*, and *utility* were related with the participant's *age*, *gender*, *general risk disposition*, and *attitude toward technology*. If one aims at studying the intention-behavior gap regarding sustainable behavior (Linder et al., 2022), one may integrate, for example, constructs such as knowledge and attitude on climate change in the research model.

## 4.4 Balancing survey length and number of participants

Determining how many topics and how many target variables should be used is not trivial and depends on many factors. An obvious consideration is the number of included topics and

2 Preliminary analysis of a still unpublished study on the perception of AI: https://osf.io/p93cy/.

dependent variables. Even if the repetitive query format facilitates efficient processing of the questionnaire (see above), both the number of topics and the number of dependent variables have an almost linear effect on the survey length.[3] Hence, the number of queried evaluation dimensions must be low. Otherwise the resulting questionnaire will be too long, resulting in reduced attention and increased dropout rates among participants. This consideration also depends on the sample and its motivation to participate: If participants are interested in the topic or are adequately compensated, more aspects can be integrated into the questionnaire. However, if participation is purely voluntary and the topic holds little interest for the participants, it is advisable to limit the number of topics and evaluation dimensions.

Defining the required sample size depends on the desired margin of error for the measurements and the empirical variance of the dependent variables used in the technology assessments. The required sample size $n$ can be calculated using the formula (Häder, 2022; Field, 2009): $n = (\frac{Z \cdot \sigma}{E})^2$ where $\sigma$ is the (unknown) standard deviation of the population, $Z$ is the critical value for the desired confidence level (for example 1.65 for a 90% confidence interval or 1.96 for a 95% confidence interval, with the latter being commonly used in the social sciences), and $E$ is the targeted margin of error in units of the dependent variable scale (e.g., 0.5 if a deviation of $\pm 0.5$ unit from the true mean is acceptable on a scale ranging from $-3$ to $+3$). The variance $\sigma^2$ can be estimated from prior research, suitable assumptions, or a pilot study. Both the desired confidence level ($Z$) and the acceptable margin of error ($E$) depend on the research goals and required precision and need to be defined by the researcher. Exploratory studies might accept higher margins of error, while confirmatory studies typically demand lower error ranges. It is important to note that if only a subset of topics is randomly sampled, this would increase the required sample size.

Based on experience, I recommend gathering at least 100 participants per topic evaluation. This sample size has yielded a margin of error of about 0.25, given the measured variance and a 95% confidence interval. By considering these factors and calculating the sample size accordingly, researchers can ensure that their findings are both statistically valid and meaningful within their research context.

## 4.5 Visualizing the outcomes

A particular advantage of this approach is that the results of the technology assessment can be clearly and accessibly presented in addition to the various possible statistical analyses.

I especially suggest the use of 2d scatter plots, which can illustrate the relationship between two dependent variables across themes (such as risk on the $x$-axis and utility on the $y$-axis), or of one dependent variable across two user groups (such as the risk assessment between laypeople on the $x$-axis and experts on the $y$-axis).

Since many possible visualizations can be made based on the number of different dependent variables or different groups of participants, one should focus on the most relevant ones. Here, of course, it is advisable to first select dimensions that are particularly relevant from the research question or a theoretical perspective (such as the aforementioned risk–benefit trade-off; even if the variable valence is used for calculations but not illustrated). Alternatively and especially for more exploratory studies, one can also display pairs of variables that have a particularly strong or weak relationship with each other. Note that readers will profit from good illustrations and clear annotations what the figure conveys. Hence, the axis, quadrants, and regression lines should be labeled clearly and readers should be guided through the interpretation of the diagram.

## 4.6 Drawbacks, challenges, and outlook

Besides advantages and insights, each method in the social sciences has its disadvantages and limitations. The following section discusses the limitations and challenges of the micro-scenario approach. Suitable alternatives are suggested afterwards.

Two (deliberate) limitations of the micro-scenarios are the brevity of the scenario narrative and the concise assessment using only a few response items. The consequence of this terseness is potentially less precise evaluations, likely contributing to greater variance in the data.

Since the scenarios cannot be presented in greater detail (compared to single scenario evaluations), the mental models of the participants—and these mental models are ultimately evaluated—can differ substantially and may be oversimplified. Of course, this is not necessarily a disadvantage if the research goal is the quantification of the affective evaluation of various topics (Finucane et al., 2000; Slovic et al., 2002). Nonetheless, possible alternatives should be considered and measures should be taken to mitigate this drawback of this approach.

If the topic evaluations are queried on single items scales, one cannot calculate reliability measures for the constructs [e.g., Cronbach's alpha ($\alpha$) or McDonald's Omega ($\omega$) as common measures for internal reliability]. Additionally, given the vast number of dependent variables collected ($n \times m$, represented by the product of the number of topics $n$ and the number of outcome variables $m$), a detailed analysis of each variable's distribution and associated characteristics (for instance, normality and unimodality) for each topic is impractical. The use of single-item scales by itself is doable, if one has the reasonable assumption that the measured construct is unidimensional, well-defined, and narrow in scope (Rammstedt and Beierlein, 2014; Fuchs and Diamantopoulos, 2009; Ang and Eisend, 2017; Woods and Hampson, 2005). In this respect, one should have sensible prior assumptions regarding the planned dependent variables or carry out accompanying studies to test these.

While the internal consistency cannot be calculated, one can calculate other reliability measures, such as the intraclass correlation coefficient (ICC). This measures if the raters (i.e., the participants of a study) agree with their ratings on each single-item scale across the different queried topics (Cicchetti, 1994). Consequently, higher ICCs would indicate a consistency in the evaluations, with some technologies or topics rated as higher and

3 For example, if the number of dependent variables is increased from three to four the expected survey duration grows by 33%.

others as lower. But although high consistency is important for the construction of a psychometric scale, it cannot be assumed for technology assessment: For example, society has no unanimous opinion on technologies such as nuclear power (Slovic, 1996) or wind power (Wolsink, 2007). In this respect, different opinions influence the measured ICC.

A limitation is that the interactions between a topic or a set of topics and the participants cannot easily be identified or interpreted. If the results suggest specific outliers or interactions, one is advised to re-evaluate the specific technologies using alternative methods for mental model extraction (such as topic-specific surveys or interviews) that allow more robust measurements in exchange for less queried topics.

When evaluating scenarios, it is essential that a good scenario description evokes a clear mental model in the participants and that they can evaluate it as accurately as possible with regard to the research question. Even more than in studies with one or a few scenarios, in the micro-scenario approach researchers must ensure that the scenarios are formulated concisely and that the response items can be clearly interpreted by the participants. Due to the breadth of topics covered in a micro-scenario study, intensive pretesting of the scenario descriptions, the evaluation metrics and the tools used is essential.

One solution to mitigate these issues could involve providing lengthier and more detailed scenario description alongside more comprehensive response items. However, maintaining the questionnaire's duration constant would necessitate a transition to a between-subjects design or the partitioning of scenarios and their evaluations across multiple studies. In an extreme scenario, a cumulative evaluation could be constructed through a meta-analysis across numerous studies with a similar structure. Such measures would undeniably enrich the validity of the results but at the cost of requiring substantially more participants and resources. Hence, this would annihilate the advantages that the micro-scenario methodology offers, such as a within-subject measurement, efficiency, and rapid data collection.

As noted earlier, studies suggest that the relationships between survey items and constructs can be distorted by lexical biases, such as word co-occurrences (Gefen and Larsen, 2017). While micro-scenarios alone won't fully resolve this issue, they can help explain and mitigate its effects. Unlike abstract or generalized survey items, micro-scenarios present specific, contextualized situations. This specificity may reduce the impact of lexical similarity, which can otherwise skew responses due to the proximity of wording rather than reflecting genuine differences in perception, particularly when comparing studies from different contexts but with the same outcome variables. By integrating multiple scenarios into a single comprehensive survey, micro-scenarios enable the evaluation of a wide range of technologies and concepts simultaneously. This approach captures more nuanced insights and reflects a broader spectrum of user experiences, reducing the reliance on potentially biased single-topic constructs. Furthermore, micro-scenarios facilitate reflexive measurement across different technologies or topics, better accounting for individual differences in technology perception. This goes beyond surface-level responses, revealing deeper patterns in how people relate to technology, thus addressing limitations in traditional methods. In summary, micro-scenarios may reduce lexical biases and enhance the robustness of technology acceptance assessments by complementing traditional methods with a more contextualized, comprehensive, and nuanced approach to understanding public perception.

## 5 Application example

To make the application and potential outcomes of this method more tangible, I will present the structure and results of an study on the acceptance of medical technology I contributed to. Detailed information on the goal of the article, its methodological approach, sample, and results can be found in the corresponding article (Brauner and Offermann, 2024). The aim of the study was to investigate how various medical technologies are assessed in terms of perceived risk and benefits, as well as a general valence evaluation. Additionally, the study sought to determine which of the two predictors—perceived risk or benefits—has a stronger influence on valence, and whether user factors affect this evaluation.

Initially, we compiled a list of 20 different medical technologies in workshops, ensuring a balance between older and newer, as well as invasive and non-invasive technologies. The technologies ranged from adhesive bandages and X-rays to mRNA vaccines. We then had these technologies evaluated by 193 participants using the assessment dimensions of perceived risk, perceived benefit, and general valence (ranging from negative to positive).

The results are 3-fold:

First, in general and across all queried technologies and participants, medical technologies are perceived as rather safe (Risk = −44.5%) and useful (Utility = 48.4%) by the participants. Similarly, the overall attributed valence—that is how positive or negative the participants evaluate the technology—is rather positive (Valence = 49.0%). Figure 5 illustrates the distribution of the evaluations.

Second, when the overall assessments of the topics were interpreted as an individual difference (Perspective 1, see Section 3), the results suggest that the valence toward medical technology is linked by individual differences, with caregiving experience and trust in physicians emerging as significant predictors.

Third, Figure 6 illustrates the risk-utility tradeoffs and the negative relationship between perceived risk and perceived benefits ($r = -0.647$, $p = 0.02$). It shows that technologies like "home emergency call button" and "plaster cast" are both highly useful and carry low perceived risk, whereas "robotic surgery" and "insulin pumps" are seen as useful but carry higher perceived risks. Finally, the novel "mRNA vaccines" are perceived as relatively high risk and low utility compared to other technologies in this study, which might reflect public skepticism or misinformation during the survey period. Furthermore, a regression analysis suggest that much of the variance in valence ($R^2 = 0.959$) is predicted by utility ($\beta = 0.886$) and to a lesser extend by the perceived risk ($\beta = -0.133$). Overall, this chart provides a visual representation of the public opinion on various medical technologies and how these are perceived in terms of their risks and benefits. It helps to identify which technologies are most favorably viewed (top-left quadrant) and which are viewed with skepticism (bottom-right quadrant). It can inform policymakers, healthcare providers, and technology developers on areas where

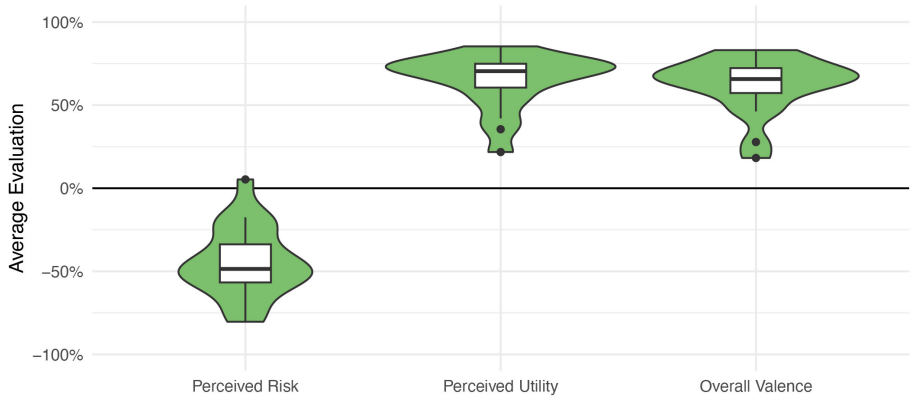



FIGURE 5
Average evaluations of 20 medical technologies by 193 participants showing that most medical technologies are seen as low risk, beneficial, and positive. Adapted from Brauner and Offermann (2024).

perceptions of risk and utility may need to be addressed, which could be crucial for adoption strategies, communication plans, and further research.

## 6 Conclusion

Overall, the presented approach enables a superordinate comparison and visualization of the acceptance and perception of a broad variety of technologies and concepts (context-specifically or cross-contextually) on different measures.

The interaction between technology and people and their values is complex and multifaceted. Some technologies can be directly compared based on public perception and mental models, particularly those within the same domain or serving similar functions. Others may require more nuanced, context-specific evaluations. This section discusses key insights, advantages, and limitations of this approach.

In general, the *approach is pragmatic* and provides an accessible comparative overview of the acceptance and perception of technologies or technology-related concepts by integrating the evaluation of *many topics* (i.e., diverse technologies in a specific or various contexts) in a single comprehensive study. This entails many advantages as it *can inform various target groups* about potentially critical issues. For example, for technology developers and researchers, this approach provides ideas and starting points to improve and develop critical technologies alongside future users' needs and perceptions. For social scientists, insights from this approach enables them to derive recommendations regarding information and (risk) communication to address future users' needs and requirements. Finally, the insights of this approach can also be used by policymakers as the basis for decision-making for governance, as it provides information about what has to be controlled better, where priorities should be set within the development and realization of innovation technologies and applications, and where citizens need more information and involvement.

Beyond the comparative overview the approach offers methodological benefits: First of all, the approach enables the transformation of the topic evaluations into *visual cognitive maps*. Herein, the different topics from the same domain can be viewed in their spatial relation to the other topics and their absolute placement. Furthermore, the relationships between the queried target variables can be statistically analyzed, for example, by interpreting their correlations, slopes, and intercepts. Various perspectives can be studied (partially based on the visualizations) within the introduced approach: A *contextual analysis* provides insights on how different topics are related to each other, and reveals potential outliers. Furthermore, the placement of the dots (as the *mean evaluations* of each topics) on the axes show how the topics are placed and perceived (e.g., rather risky or not). The *dispersion*, that is, the distribution of the dots across the scales, indicates the consistency of the evaluations and shows whether they represent uniform or rather diverse attributions. Further, *correlations* between the attributions can be analyzed and show how strong different evaluations are related across the topics. Additionally, different *intercepts on the axes* and thus the position of the topics can also be analyzed and interpreted. If three or more variables are evaluated per topic and one is a dedicated target variable, the *degree of explained variance* can be interpreted by means of regression analyses, to inform how uniform the topic evaluation is across all topics. Regression analyses also inform which factors have the *strongest influence* on the target variables (such as valence). These results can then be used to, for example, derive adequate and tailored communication strategies. Finally, as with other approaches, the overall evaluations per participants can be *linked to other responses from the participant*, such as their demographics, attitudes, beliefs, or reported behaviors. In this regard, the introduced mapping and visualizations of the evaluations can also be realized to compare different sub-samples depending on specific variables (e.g., age groups, low vs. high technology expertise).

In addition, the article provides *practical tools* in terms of specific recommendations and R code alongside the

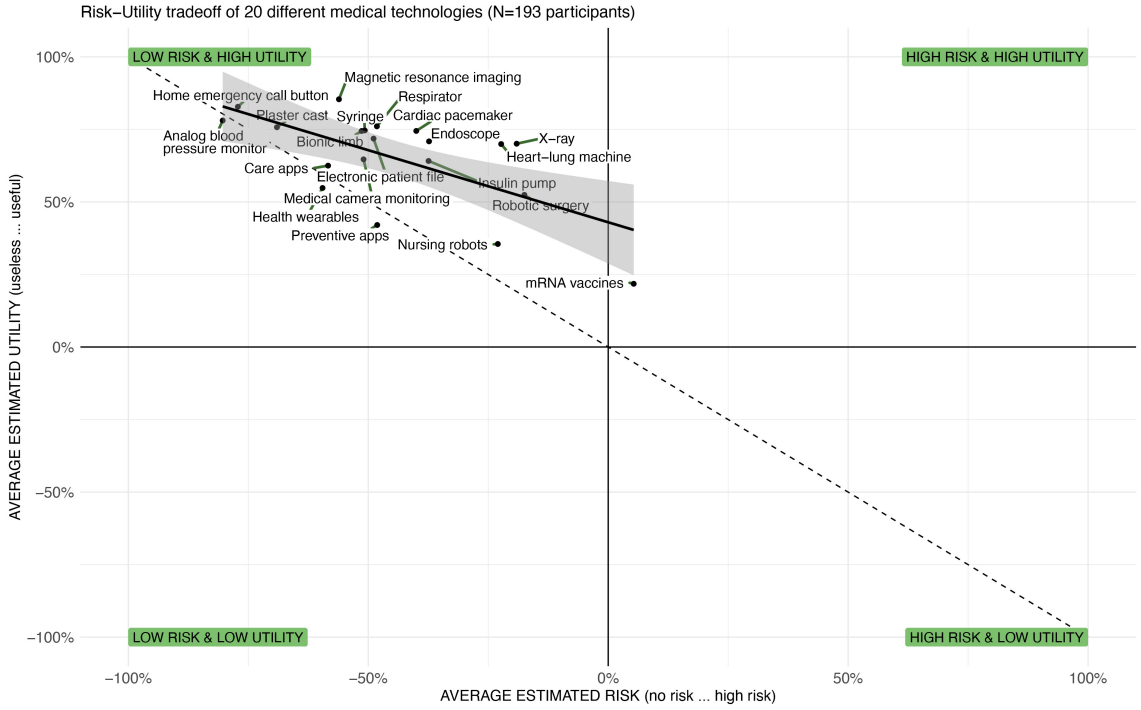



FIGURE 6
Illustration of the risk-utility tradeoffs across 20 medical technologies showing that most are evaluated as rather safe and useful and a strong correlation between both measures (adapted from Brauner and Offermann, 2024).

methodological concept, which will help easily use and directly apply the presented approach in future research.

Summarizing the methodological advancements of the micro-scenario approach, the dual complementary perspectives offer three significant benefits. First, they facilitate the modeling of individual differences through reflexive measurement across various technologies or topics. Second, they provide valuable insights for developers, researchers, and policymakers by analyzing the spatial positioning of the topics to identifying critical issues in technology perception. Third, this enables the identification of acceptance-relevant factors crucial for tailoring technology to better meet human needs.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author. An executable R notebook that extends the example code provided in this article is publicly available https://github.com/braunerphilipp/MappingAcceptance. All data used for generating the example is available at the repository.

## Author contributions

PB: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612. Open access funding provided by the Open Access Publishing Fund of RWTH Aachen University.

## Acknowledgments

This approach evolved over time and through several research projects. I would like to thank all those who have directly or indirectly, consciously or unconsciously, inspired me to take a closer look at this approach and who have given me the opportunity to apply this approach in various contexts. In particular, I would like to thank: Julia Offermann, for invaluable discussions about this approach and so much encouragement and constructive comments during the final meters of the manuscript. Martina Ziefle for igniting scientific curiosity and motivating me to embark on a journey of boundless creativity and exploration. Ralf Philipsen, without whom the very first study with that approach would never have happened, as we developed the crazy idea to explore the

benefits of barriers of using "side-by-side" questions in Limesurvey. Julian Hildebrandt for in-depth discussions on the approach and for validating the accompanying code. Tim Schmeckel for feedback on the draft of this article. Felix Glawe, Luca Liehner, and Luisa Vervier for working on a study that took this concept to another level. Of course, I would also like to thank my two referees and the editors. They were very accurate in their identification of weaknesses and ambiguities, and their very constructive feedback really contributed to the strengthening of the article. I express my gratitude to my son Oskar, who—while teething and sleeping on my belly at night—allowed me to write substantial portions of this article. I utilized Large Language Models (LLMs), specifically OpenAI's GPT-3.5 and GPT-4o, for assistance in editing and R coding. For writing assistance, typical prompts included requests such as, "I am a scientist writing an academic article. Can you edit the following paragraph? Please explain the changes you made and the rationale behind them." For coding support, prompts included, "I am coding in R and have two data frames, A and B. How do I merge these using the unique ID in `tidyverse` syntax?" Importantly, LLMs were not used to generate (or "hallucinate") any original content for the manuscript.

"Your methods turn out to be ideal. Go ahead."—Quote from one of my last fortune cookies. No scientific method of the social sciences alone will fully answer all of our questions. I hope that this method provides a fresh perspective on exciting and relevant questions.

If you have carried out a study based on this method, please let me know with the research context and the investigated variables and I will document it on the project's page (https://github.com/braunerphilipp/MappingAcceptance).

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Agogo, D., and Hess, T. J. (2018). "How does tech make you feel?" a review and examination of negative affective responses to technology use. *Eur. J. Inf. Syst.* 27, 570–599. doi: 10.1080/0960085X.2018.1435230

Al-Emran, M., Mezhuyev, V., and Kamaludin, A. (2018). Technology acceptance model in m-learning context: a systematic review. *Comp. Educ.* 125, 389–412. doi: 10.1016/j.compedu.2018.06.008

Ang, L., and Eisend, M. (2017). Single versus multiple measurement of attitudes: a meta-analysis of advertising studies validates the single-item measure approach. *J. Advert. Res.* 58, 218–227. doi: 10.2501/JAR-2017-001

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometr. Bull.* 2, 47–53. doi: 10.2307/3002000

Bertogg, A., Imdorf, C., Hyggen, C., Parsanaglou, D., and Stoilova, R. (2020). Gender discrimination in the hiring of skilled professionals in two male-dominated occupational fields: a factorial survey experiment with real-world vacancies and recruiters in four European countries. *Soziol. Sozialpsychol.* 72, 261–289. doi: 10.1007/s11577-020-00671-6

Brauner, P., Hick, A., Philipsen, R., and Ziefle, M. (2023). What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. *Front. Comp. Sci.* 5:1113903. doi: 10.3389/fcomp.2023.1113903

Brauner, P., Liehner, G. L., Vervier, L., and Ziee, M. (2024). Modelling the risk-utility tradeoff in public perceptions of artificial intelligence. *OSF.io*. Available at: https://osf.io/p93cy

Brauner, P., and Offermann, J. (2024). Perceived risk-utility tradeoffs of medical technology: a visual mapping. *SocArXiv*. doi: 10.31235/osf.io/cfvq9

Breakwell, G. M. (2001). Mental models and social representations of hazards: the significance of identity processes. *J. Risk Res.* 4, 341–351. doi: 10.1080/13669870110062730

Buse, R. P., Sadowski, C., and Weimer, W. (2011). "Benefits and barriers of user evaluation in software engineering research," in *Proceedings of the 2011 ACM International Conference on Object Oriented Programming Systems Languages and Applications* (New York, NY: ACM), 643–656. doi: 10.1145/2076021.2048117

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* 6, 284–290. doi: 10.1037/1040-3590.6.4.284

Courage, C., and Baxter, K. (2005). *Understanding Your Users - A Practical Guide to User Requirements Methods, Tools & Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Craik, K. J. W. (1943). *The Nature of Explanation*. New York, NY: Cambridge University Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi: 10.2307/249008

Döring, N. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin; Heidelberg: Springer.

Eisenstein, E. (1980). *The Printing Press as an Agent of Change: Communications and Cultural, Transformations in Early-Modern Europe*. New York, NY: Cambridge University Press.

Engels, F. (1845). *The Situation of the Working Class in England*. Leipzig: Druck und Verlag Otto Wigand.

Field, A. (2009). *Discovering Statistics Using SPSS, 3rd Edn*. Thousand Oaks, CA: Sage Publications.

Finucane, M. L., Alhakami, A., Slovic, P., and Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *J. Behav. Decis. Mak.* 13, 1–17. doi: 10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Fischhoff, B. (2015). The realities of risk-cost-benefit analysis. *Science* 350:eaaa6516. doi: 10.1126/science.aaa6516

Fuchs, C., and Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research - conceptual issues and application guidelines. *Die Betriebswirtschaft* 9, 195–210.

Gefen, D., and Larsen, K. (2017). Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. *J. Assoc. Inf. Syst.* 18, 727–757. doi: 10.17705/1jais.00469

Gigerenzer, G., and Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Top. Cogn. Sci.* 1, 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gilovich, T., Griffi, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment, 8th Edn*. New York, NY: Cambridge University Press.

Groves, R. M., Fowler, F. J., Couper, M. K., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology. 2nd Edn*. Hoboken, NJ: Wiley.

Grunwald, A. (2009). "Technology assessment: concepts and methods," in *Philosophy of Technology and Engineering Sciences* (Elsevier), 1103–1146.

Guston, D. H., and Sarewitz, D. (2002). Real-time technology assessment. *Technol. Soc.* 4, 93–109. doi: 10.1016/S0160-791X(01)00047-1

Häder, M. (2022). *Empirical Social Research: An Introduction*. Germany: Springer Wiesbaden.

Johnson-Laird, P. N. (2010). Mental Models and Human Reasoning. *Proc. Natl. Acad. Sci. U. S. A*. 107, 18243–18250. doi: 10.1073/pnas.1012933107

Jones, N. A., Ross, H., Lynam, T., Perez, P., and Leitch, A. (2011). Mental models: an interdisciplinary synthesis of theory and methods. *Ecol. Soc*. 16:146. doi: 10.5751/ES-03802-160146

Kim, H.-W., Chan, H. C., and Gupta, S. (2007). Value-based adoption of mobile internet: an empirical investigation. *Decis. Support Syst*. 43, 111–126. doi: 10.1016/j.dss.2005.05.009

Kosow, H., and Gaßner, R. (2008). *Methods of Future and Scenario Analysis: Overview, Assessment, and Selection Criteria, Vol. 39*. Bonn: German Development Institute.

Kübler, D., Schmid, J., and Stüber, R. (2018). Gender discrimination in hiring across occupations: a nationally-representative vignette study. *Labour Econ*. 55, 215–229. doi: 10.1016/j.labeco.2018.10.002

Kvavadze, E., Bar-Yosef, O., Belfer-Cohen, A., Boaretto, E., Jakeli, N., Matskevich, Z., et al. (2009). 30,000-year-old wild flax fibers. *Science* 325:1359. doi: 10.1126/science.1175404

Lidynia, C., Brauner, P., and Ziefle, M. (2017). *A Step in the Right Direction – Understanding Privacy Concerns and Perceived Sensitivity of Fitness Trackers*. Cham: Springer International Publishing, 42–53.

Linder, N., Giusti, M., Samuelsson, K., and Stephen, B. (2022). Pro-environmental habits: an underexplored research agenda in sustainability science. *Ambio* 51, 546–556. doi: 10.1007/s13280-021-01619-6

Luce, R. D., and Tukey, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *J. Math. Psychol*. 1, 1–27. doi: 10.1016/0022-2496(64)90015-X

Lucke, D. (1995). *Akzeptanz*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Marangunić, N., and Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Univ. Access Inf. Soc*. 14, 81–95. doi: 10.1007/s10209-014-0348-1

Marikyan, D., Papagiannidis, S., and Stewart, G. (2023). Technology acceptance research: meta-analysis. *J. Inf. Sci*. doi: 10.1177/01655515231191177

Messick, S. J. (1957). Metric properties of the semantic differential. *Educ. Psychol. Meas*. 17, 200–206. doi: 10.1177/001316445701700203

Montgomery, D. C. (2019). *Design and Analysis of Experiments. 10th Edn*. Hoboken, NJ: Wiley.

Niinimäki, K., Peters, G., Dahlbo, H., Perry, P., Rissanen, T., and Gwilt, A. (2020). The environmental price of fast fashion. *Nat. Rev. Earth Environ*. 1, 189–200. doi: 10.1038/s43017-020-0039-9

Offermann-van Heek, J., and Ziefle, M. (2019). Nothing else matters! Trade-offs between perceived benefits and barriers of AAL technology usage. *Front. Public Health* 7:134. doi: 10.3389/fpubh.2019.00134

Protzko, J., and Schooler, J. W. (2023). What i didn't grow up with is dangerous: personal experience with a new technology or societal change reduces the belief that it corrupts youth. *Front. Psychol*. 14:1017313. doi: 10.3389/fpsyg.2023.1017313

Rahimi, B., Nadri, H., Lotfnezhad Afshar, H., and Timpkas, T. (2018). A systematic review of the technology acceptance model in health informatics. *Appl. Clin. Inform*. 9, 604–634. doi: 10.1055/s-0038-1668091

Rammstedt, B., and Beierlein, C. (2014). Can't we make it any shorter?: the limits of personality assessment and ways to overcome them. *J. Ind. Differ*. 35, 212–220. doi: 10.1027/1614-0001/a000141

Recchia, G., Freeman, A. L. J., and Spiegelhalter, D. (2021). How well did experts and laypeople forecast the size of the Covid-19 pandemic? *PLoS ONE* 16:e0250935. doi: 10.1371/journal.pone.0250935

Revelle, W., and Garner, K. M. (2024). "Measurement: reliability, construct validation, and scale construction," in *Handbook of Research Methods in Social and Personality Psychology, 3rd Edn*, eds. H. T. Reis, T. Wesrt, and C. M. Judd (Cambridge: Cambridge University Press).

Ries, A., and Trout, J. (2001). *Positioning: The Battle for Your Mind*. New York, NY: McGraw-Hill Education.

Rogers, E. M., Singhal, A., and Quinlan, M. M. (2019). "Diffusion of innovations," in *An Integrated Approach to Communication Theory and Research* (London: Routledge), 415–434. doi: 10.4324/9780203710753-35

Slovic, P. (1987). Perception of risk. *Science* 236, 280–285. doi: 10.1126/science.3563507

Slovic, P. (1996). Perception of risk from radiation. *Radiat. Prot. Dosimet*. 68, 165–180. doi: 10.1093/oxfordjournals.rpd.a031860

Slovic, P., Finucane, M., Peters, E., and MacGregor, D. G. (2002). "The affect heuristic," in *Heuristics and Biases: The Psychology of Intuitive Judgment*, eds. D. Kahneman, D. Griffin, and T. Gilovich (Cambridge: Cambridge University Press), 397–420.

Steinberg, S. (1974). *Fife Hundred Years of Printing*. Harmondsworth: Penguin.

Tolsdorf, J., Reinhardt, D., and Iacono, L. L. (2022). Employees' privacy perceptions: exploring the dimensionality and antecedents of personal data sensitivity and willingness to disclose. *Proc. Privacy Enhanc. Technol*. 2022, 68–94. doi: 10.2478/popets-2022-0036

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tran, T. A., and Daim, T. (2008). A taxonomic review of methods and tools applied in technology assessment. *Technol. Forecast. Soc. Change* 75, 1396–1405. doi: 10.1016/j.techfore.2008.04.004

Venkatesh, V., Thong, J. Y. L., and Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Q*. 36, 157–178. doi: 10.2307/41410412

Verhagen, T., Hooff, B., and Meents, S. (2018). Toward a better use of the semantic differential in is research: an integrative framework of suggested action. *J. Assoc. Inf. Syst*. 16, 108–143. doi: 10.17705/1jais.00388

Watt, J. (1769). *New Invented Method of Lessening the Consumption of Steam and Fuel in Fire Engines, British Patent No. 913*.

Wolsink, M. (2007). Wind power implementation: the nature of public attitudes: equity and fairness instead of 'backyard motives'. *Renew. Sustain. Energy Rev*. 11, 1188–1207. doi: 10.1016/j.rser.2005.10.005

Woods, S. A., and Hampson, S. E. (2005). Measuring the big five with single items using a bipolar response scale. *Eur. J. Pers*. 19, 373–390. doi: 10.1002/per.542

Zhang, P., Li, N., and Sun, H. (2006). "Affective quality and cognitive absorption: extending technology acceptance research," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, *Vol. 8* (Piscataway, NJ: IEEE), 207a. doi: 10.1109/HICSS.2006.39

# Appendix: practical tools: implementing and analyzing micro scenarios

This appendix provides practical tips for implementing micro scenarios in surveys and analyzing the resulting data. An executable R notebook[4] offers a comprehensive example, including code for data transformation, analysis, and visualization.

## Implementing micro scenarios in survey tools

Many survey tools simplify the creation, data collection, and analysis of online questionnaires, reducing the need for manual input.

For example, the *side-by-side* question format (available in tools like Limesurvey and Qualtrics) displays topics and their response items as rows in a table. While easy to process, this format requires all items to be displayed on the same page, which may overwhelm participants or be difficult to view on small screens.

Some tools like Qualtrics offer advanced options such as *Loop & Merge*, which generates repeating blocks based on a data table (e.g., topic titles and descriptions). The tool iterates through all or a subset of topics, presenting them with consistent formatting. Survey data is stored in a structured format, with response variables named systematically (e.g., `aN_matrix_M`, where `N` is the topic number and `M` the dimension).

## Data analysis

Standardized variable names, like those generated by *Loop & Merge*, allow for systematic and automated data transformation. Below, I provide R code examples using the `tidyverse` package (https://www.tidyverse.org/). Other software can also be used.

### Rearranging survey data from wide to long format

Survey data must be reshaped from wide format (one row per participant, as in Figure 2) to long format (one value per row for each participant, topic, and evaluation dimension). Listing 1 demonstrates this transformation using `pivot_longer`. Additionally, survey responses (e.g., 1–7 scales) are rescaled to a percentage format ranging from −100% to +100%. Other variables, such as demographics, are neglected but will be added at a later stage.

Listing 1 Convert survey data to long format (one row per observation.

```
long <- surveydata %>%
dplyr::select(id, matches("a\\d+
\\_matrix\\_\\d+")) %>%
tidyr::pivot_longer(
cols = matches("a\\d+\\_matrix\\_\\d+"),
names_to = c("question", "dimension"),
names_pattern = "(.*)_matrix_(.*)",
# Separate topic and evaluation
values_to = "value",
values_drop_na = FALSE) %>%
dplyr::mutate( dimension = as.numeric
(dimension) ) %>% # readable dims
dplyr::mutate( dimension = DIMENSIONS
[dimension]) %>%
dplyr::mutate( value = -(((value - 1)/3)
- 1)) # rescale [1..7] to [-100%..100%]
```

### Calculating grand means for dimensions

In Listing 2, the grand mean for each assessment dimension is calculated across all topics and participants.

Listing 2 Calculate grand mean for each assessment dimension.

```
byDimension <- long %>%
dplyr::group_by( dimension ) %>%
dplyr::summarise( mean = mean(value,
na.rm = TRUE),
sd = sd(value, na.rm = TRUE), .groups="drop")
```

### Research perspective 1: calculate average topic evaluations as individual differences

Listing 3 shows how to pivot the data back to wide format and calculate the average topic evaluations for each participant. After pivoting, participants' topic evaluations are aggregated (e.g., by mean or median). The resulting data has one row per participant and columns for the average evaluation. These results can be merged with original survey responses using `left_join` to explore relationships with other variables (see Section 4.3).

Listing 3 Calculate average topic evaluations for each participant.

```
byParticipant <- long %>%
tidyr::pivot_wider(
names_from = dimension,
values_from = value) %>%
dplyr::group_by(id) %>%
dplyr::summarize(
across(
all_of( DIMENSIONS ), # Select evaluation
dimensions
list( mean = ~mean(., na.rm = TRUE),
median = ~median(., na.rm = TRUE),
sd = ~sd(., na.rm = TRUE)),
.names = "{.col}_{.fn}" # Define schema for
column names
), .groups="drop"
) %>%
dplyr::left_join(surveydata, by="id")
```

4 https://github.com/braunerphilipp/MappingAcceptance

## Research perspective 2: calculate average topic evaluations

Listing 4 shows how to calculate the average assessments for each topic by summarizing data using the arithmetic mean and standard deviation. As shown in Figure 3, the data now contains one row per topic with two columns (mean and SD) for each dimension. This topic-related data can now be analysed or visualised.

Listing 4 Calculate average evaluations for each topic.

```
byTopic <- long %>%
tidyr::pivot_wider(
names_from = dimension,
values_from = value) %>%
dplyr::group_by( question ) %>%
dplyr::summarize(
across(
all_of( DIMENSIONS ), # Select evaluation
dimensions
list(mean = ~mean(., na.rm = TRUE),
sd = ~sd(., na.rm = TRUE)),
.names = "{.col}_{.fn}" # Define schema for
 column names
), .groups="drop")
```

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 13 November 2024
DOI 10.3389/fpsyg.2024.1410047



OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
Kenneth Gergen,
Swarthmore College, United States
Roland Mayrhofer,
University of Regensburg, Germany

*CORRESPONDENCE
Julia Scholz
✉ jscholz@uni-wuppertal.de

RECEIVED 31 March 2024
ACCEPTED 22 October 2024
PUBLISHED 13 November 2024

CITATION
Scholz J (2024) Agential realism as an alternative philosophy of science perspective for quantitative psychology.
*Front. Psychol.* 15:1410047.
doi: 10.3389/fpsyg.2024.1410047



# Agential realism as an alternative philosophy of science perspective for quantitative psychology

Julia Scholz*

Institute of Psychology, University of Wuppertal, Wuppertal, Germany

This paper introduces Karen Barad's philosophical framework of agential realism as an alternative philosophy of science perspective for quantitative psychology and measurement. Agential realism offers a rethinking of the research object, measurement process and outcome, causality, and the researcher's responsibility by proposing an ethico-epistem-ontological understanding of material-discursive practices that co-construct our world. The contemporary, canonical underlying philosophy of science perspective of quantitative psychology entails entity realism, a difference between ontic existence and epistemic approaches, complete causality, and determinism. Consequently, the researcher has no responsibility for the characteristics of a research object. The paper introduces agential realism and its assumptions about rejecting entity realism but a particular understanding of phenomena, the entanglement of ontic existence and epistemic approaches, and the researcher's role in co-creating an outcome. A reworking of the concept of causality implies newly emerging possibilities for realizations. Subsequently, the paper addresses four consequences of applying agential realism in quantitative psychology. (1) If there is indeterminacy in every phenomenon, researchers do not search for one true score but assume a realization potential, which has implications for comparisons and replications. (2) If configurations are part of things-in-phenomena, then context does not work as a third variable; instead, all 'parts' are co-creators. This entanglement must be considered in replications instead of trying to eliminate its impact. (3) Agential realism encompasses the researchers' responsibility to justify decisions made in a research project and to clarify ethics. (4) Overall, agential realism alters the research endeavor by asking new questions and interpreting research outcomes differently. Further directions point towards concrete tasks like methodological questions and the necessity within psychology to elaborate further on the conceptualizations initiated by Barad.

KEYWORDS

**agential realism, philosophy of science, intra-action, phenomena, agential cut, replication, ethico-onto-epistemology, realization potential**

## 1 Introduction

Psychological science faces, once again, discussions about its knowledge acquisition. The discussions should be seen as a sign of quality: science is open to questioning. Some 'crises' in the field forced psychologists, alongside colleagues from other disciplines, to reconsider what their knowledge represents. Psychologists' knowledge is usually aimed at describing, understanding, explaining, and sometimes changing human thought, feeling, experience, and behavior. Besides these knowledge fields, psychology is also concerned about how psychological knowledge is gained. Besides previous debates about experimental research logic (*cf.* Gergen, 2001), recently, epistemological, conceptual, and methodological challenges in psychological science practices are again discussed (*cf.* Hanfstingl et al., 2023). For instance, the Open Science Collaboration (2015)

identified a low replication rate, but Gilbert et al. (2016) accused the project of underestimating it. In the face of such discussions—Nosek et al. (2022) named the 2010s 'a decade of active confrontation'—many very sophisticated articles analyzed statistical and methodological problems, and researchers devised various solutions to increase the replicability of experimental results. However, methodological procedure is not the focus of this paper. I will not discuss improvements in accomplishing and processing the contemporary standard quantitative method. Instead, I will discuss a shift in the philosophy of science perspective for quantitative psychology, and consider its consequences, also for replication. First, I will look at the underlying basis of contemporary quantitative psychology, and then I will propose an alternative: Karen Barad's agential realism.

Agential realism was extensively adapted in a wide range of fields. For example, Barad's article 'Posthumanist Performativity' (2003) has already been cited more than 10.000 times. The book 'Meeting the Universe Halfway' (2007) more than 20.000 times. Hollin et al. (2017) give a little peek at Barad's reception regarding content. However, there is only sparse reception within psychology, and if so, then primarily within qualitative approaches (e.g., Brown, 2020; Gemignani et al., 2023). Mauthner (2024) discusses broad changes within a research logic if we[1] take an agential realist perspective but also brings qualitative methods to the fore. Besides my own work (Scholz, 2013, 2018), it was, for instance, Shotter (2014a) who encouraged psychologists to take Barad's perspectives as a matter of principle. Next to a few discussions in the journal 'Theory & Psychology' (Højgaard and Søndergaard, 2011; Shotter, 2014b; Tobias-Renstrøm and Køppe, 2020), for example, Letiche et al. (2023) discussed agential realism for experiments and called for reworking of quality criteria of research. However, this was centered on 'accounting' and not directly about psychological experiments. Agential realism is hardly ever applied to quantitative research in psychology or questions of replicability. I will get back to this below, but first, I will look at the underlying philosophy of science perspective of quantitative psychology.

## 2 The underlying philosophy of science perspective of quantitative psychology

Every working paradigm has its foundational logic about why somebody is doing something, such as researchers having a reason to do science this way or that way. The starting points of every paradigm are pre-assumptions about the world from which methods are deduced. On the other hand, somebody who uses methods has pre-assumptions about the world in which the specific method makes sense. Psychology researchers typically do not explicate their foundational pre-assumptions in a research article, but these can mostly be read from the researcher's proceeding or wording. Other pre-assumptions lead to a different proceeding or other wording. Regarding research logic, Popper's 'logic of scientific discovery' is still widely used, though further developed and enhanced. I will mention where this logic plays a role in my argument but will not summarize it entirely. Instead, I will concentrate on the points still used in quantitative psychology but contrasted with the proposed alternative of agential realism. Table 1 offers shortened descriptions of conceptualizations from each perspective in a comparative manner.

In this text, I will, as Uher (2022) also urges, be sensitive to the distinction between 'psychology' and 'psyche', although many psychological texts do not make a clear distinction and use 'psychology' when referring to 'psyche.' However, since I will also address the approach of the discipline of psychology, I need to be clear in sentences whether I am talking about the discipline or the human psyche. Likewise, I will differentiate between ontic and ontological, as well as between epistemic and epistemological—find an overview of such differentiations in Table 2.

### 2.1 Entity realism

To start, I will examine the understanding of the constitution of the research objects within quantitative psychology. By 'objects' (Table 1), I mean the subject matter of psychology. It is that what is described with nouns in that discipline. These nouns can refer to physical things like 'neuron' or 'lens' but also to concepts like 'self-confidence' or 'sensibility', concrete experiences like 'fear', and broader categories like 'behavior' or 'feeling.' To compare philosophy of science perspectives, I will also use 'entity' (Table 1) for such a subject matter. Neurons, sensibility, fear, or behavior are all 'entities' and 'research objects' of the discipline of psychology.

Contemporary quantitative psychology comprises realism toward these research objects in that they are preexisting objects (i.e., individually determinate bounded entities) with inherent properties. This entity realism is one central assumption of the classic realist philosophy of science perspective. Dienes (2008) states some differing positions within psychology but closes that scientists need real entities to maintain a 'subject matter.' This aligns with Popper's (2002) perspective, which puts a realist position not as a requirement for the 'logic of scientific discovery' but as the background in which the pursuit of truth gains meaning. Also, Herzog (2012) states that scientists classify themselves as belonging to what they call materialism or physicalism in a classical realist way. Some psychologists explicitly state that the research objects they investigate are not merely auxiliary constructs in an instrumentalist way but are ontologically (Table 2) understood as 'real' (Table 1). "Psychologists (…) also generally believe in the reality of the domain of their subjects—of mind, and brains, thoughts, images, networks, social pressures, social identities, psychological contexts and so on" (Dienes, 2008, p. 28). It is clear that psychological objects need not be physical objects (e.g., like a neuron) but can be a process, a state, a feeling, or the like. Uher (2021) also resumes that it is widespread for psychologists to ascribe an ontic status to constructs, which is entity realism. In a hypothesis like 'increasing empathy reduces racial bias', the constructs 'empathy' and 'racial bias' are assumed to exist before the researcher enters the stage. Therefore, I also use the terms 'entity' and 'objects' in this psychological realm for occurrences like 'behavior' or 'emotion.' The critical point is the philosophical pre-assumption about the occurrences that a discipline investigates, which can be physical objects in physics but might be behavior in psychology. In the following explanations,

1 I use 'we' in this article when discussing philosophy of science perspectives because that is the topic I offer in this paper. I use 'researchers' when discussing concrete consequences of such perspectives for research practices.

TABLE 1 Comparison of concepts between the contemporary, canonical psychological, and the agential realist perspective.

| Contemporary, canonical psychological perspective | Concept | Agential realist perspective |
|---|---|---|
| (No application) | Agential | The adjective is used to indicate that a correspondent referent ('cut', 'realism', or 'separability') does not exist *per se* but that an agency brings this referent into being |
| Bounded entity that is built of components and that can measure and/or manipulate something | Apparatus | Enacts intra-actions and agential cuts within the phenomenon; is itself entangled in material-discursive configurations |
| A vector of influence is transported from one process, state, or object to another process, state, or object | Causal | Intra-actions can enact a causal structure; causality is entangled with conditionality |
| bounded, determined occurrence with pre-existing properties/features (independent from a measurement process); can also be composed of several smaller components | Entity / object / relatum (here also: that subject matter of psychology which is referred to by nouns, like 'neuron', 'sensibility', 'fear' or 'behavior') | always unbounded as entity-within-phenomena or thing-within-phenomena or relatum-within-relations; does not exist without material-discursive configurations that, through intra-actions, enact this occurrence |
| Inter means between; interaction as an action between separate entities; separate entities influence each other or one entity influences the other | Interaction vs. intra-action | Intra means within or inside; intra-action as an action within an entanglement; intra-acting relata are not separate entities but relata-within-relations |
| Refers to having information about objective facts; if we 'know' something is an epistemological question | Knowing, knowledge | Knowing and being are mutually implicated: if we 'know' something is an onto-epistemological question |
| Material and discursive, or social, influences can impact a process, state, or object causally | Material-discursive configurations | Entangled configurations that enact intra-actions and set agential cuts; they situatedly co-create what exists |
| A (numerical) representation of a 'real-world occurrence' | Measurement outcome | An occurrence enacted from intra-actions of material-discursive configurations |
| The structurally identical assignment of a numerical relative to an empirical one; an epistemic activity | Measurement process | An intra-action that enacts agential cuts; carves out one of several possibilities; co-creates characteristics of the research object |
| Independent of an onlooker, subject, or researcher | Objective | Accountive of the constitutive material-discursive configurations |
| A complex unit that entails different components and an inner structure | Phenomenon | An entangled non-bounded occurrence that can enact further agential cuts |
| Real is existent independent from an onlooker; reality refers to an objective world that exists independent of perceptions, beliefs, or thoughts | Real, reality | Something is real if there is a locally and temporarily shared 'experience' of intra-actions and cuts; not necessarily a human 'experience', can, for example, also be 'experienced' by physical radiation; reality is always a situated reality, dependent on material-discursive configurations; |
| Discovers what was already there | Research, science | Understands situated possibilities within material-discursive practices |

psychological entities and objects are not particles but 'thoughts, images, networks' and so on.

Researchers might acknowledge that such objects have developed over time (e.g., in the history of humanity), and they might call them 'phenomena' (instead of objects) to express that several components belong to such an item. However, in the moment of theorizing about and experimenting with such concepts like 'empathy', they assume each one is a particular set of parts (e.g., two persons) with their qualities (e.g., a feeling, a thought, a motive, an ability) and their relations to one another (e.g., one person has the ability to understand and possibly share a perspective or feeling of the other person). All such entities are assumed to be 'real' in that they exist individually and independently of any onlooker (Table 1).

## 2.2 A difference between ontic existence and epistemic approaches and the understanding of measurement

One basis of Popper's (e.g., 2002) idea of scientific progress, which is still the foundation for quantitative psychology, is the *differentiation* between ontic existence and epistemic approaches to the existences. In this perspective, 'ontic existence' refers to what exists as concrete and factual nature of something and 'epistemic approach' refers to the tools we use to try to gain knowledge (Table 1) of the nature of our studied research objects. This differentiation is needed to assume that the ontic state of an object at any given time and place has a factual nature independent of researchers' attempts to gain knowledge of such a factual nature. As Popper pointed out, the aim of science is (through falsification) to gain better and better descriptions and explanations of the (classic realist) objects and of the operating causal (Table 1) chains. Accordingly, our knowledge should grow in representations—as accurate as possible—of the real object or property of anything. In this perspective, the factual property or state of anything exists independent from our epistemic approaches but those approaches can be more or less suitable to deliver good representations of the factual property. Accordingly, researchers can have varying degrees of optimism about how close they might come to an 'as accurate as possible' or a 'true' representation of their research objects. However, suppose they conclude that the representations are too poor. In that case, this is

TABLE 2 Vocabulary differentiation in this text.

| | |
|---|---|
| Epistemic | Refers to anything related to knowledge |
| Epistemological | Refers to anything related to the study or theory of knowledge |
| Ontic | Refers to anything related to being |
| Ontological | Refers to anything related to the study or theory of being |
| Psychic | Refers to anything related to the human psyche |
| Psychological | Refers to anything related to the study or theory of the human psyche |
| Onto-epistemical | Refers to anything related to the entanglement of being and knowing in the world |
| Onto-epistemological | Refers to anything related to the study or theory of the entanglement of being and knowing in the world |

attributed to epistemic reasons: the method was unsuitable, or flaws within the research design disturbed the measuring. The basic assumption of quantitative research is that, in principle, the research activity does not change the research object. If a research outcome does indicate a change in the factual nature, this must be an error. Such an indication can only demonstrate that the epistemic approach was unsuitable because ontic existence is understood as independent from epistemic approaches to it.

This differentiation between ontic existence and the epistemic approach guides the understanding of the measurement process and outcome (Table 1). A short look at a classical definition reveals a typical imprecision concerning the philosophy of science perspectives: All students of psychology learn a variation of this definition: "Measurement—a central epistemic activity in science—relates a number and a quantity in an effort to estimate the magnitude of that quantity" (Trout, 2001, p. 265). However, Trout continues: "A quantity is typically a property of a physical configuration, such as length or weight, and determines a function that applies to a domain or class of objects. At this high level of abstraction, the description of the purpose and relation of measurement is metaphysically neutral, leaving *open* the question of whether the domain is observable (empirical) or unobservable (non-empirical)" (Trout, 2001, S. 265, *my emphasis*). Here, Trout discusses 'ontological' and 'epistemological' questions because they are posed within the reasoning about science (the suffix "-logy" indicates it is about the study of anything, see Table 2). If a philosopher of science discusses that an object has a property, this is an 'ontological' question about the 'ontic' state of something. Trout claims that this definition of measurement is metaphysically neutral and does not imply realism or any other perspective. Yet we have to assert that Trout's description is *not* metaphysically neutral because here, only the *epistemological* question of whether the domain is observable or unobservable is still 'open.' The *ontological* question of assuming that an entity has a pre-existing property is answered by this definition, therefore not an open question, and this reveals a classic realist philosophy of science position—that is, entity realism. Appropriate to the aims and logic of doing research in contemporary, canonical ways, psychological measuring is commonly understood as the 'epistemic activity' (see Trout above) of trying to arrive at an 'as correct as possible truth' about a quantity. Moreover, if researchers were to detect 'problems' with their measurement process, they would engage in overcoming these 'problems' and gaining a 'better measurement process', meaning that the outcomes represent 'more correctly' the true nature of the measured entity.

I conclude that most quantitative psychologists today still follow a strict entity realism ontologically speaking and that they understand the measurement process (Table 1) as an epistemic endeavor. They assume that their research objects have factual properties independent of any onlooker and that approaches to gain knowledge of these properties are more or less suitable to arrive at an as correct as possible representation of the factual property. They might be differently optimistic about what we can 'learn' about a property, a system's state, its components, and perhaps the future, but always for epistemical reasons. The research objects are understood as having their nature, shape, quality, or property *per se*, and the task of science is to measure these as correctly as possible.

## 2.3 Responsibility of the researcher

Within the reasoning and as a logical consequence of this philosophy of science perspective—involving entity realism and the assumed difference between ontic existence and epistemic approach—scientists are not responsible for the characteristics of the research object. Researchers assume they only 'discover' what is already 'out there'; they do not think they create entities—otherwise, it would be 'flawed' science. In this reasoning, researchers must ensure that the epistemic approach approximates the pre-existing entity as much as possible and as unbiased as possible. The conventional responsibility of researchers includes doing science as 'objectively' (Table 1) and 'neutrally' as possible to find the real characteristics of the investigated property.[2]

Why it is a crisis moment when replications fail is self-explanatory. Researchers hope they have found representations of real entities, relationships, and influences that are as correct as possible. The ability to replicate experiments means support for the claim that one has discovered an objective representation. Ideally, the results can also be measured by anybody else. If researchers cannot replicate a finding, it suggests that the previous finding was wrong. In section 4, I will argue that we will have a different view on replications and some different ideas of their 'problems' and 'cures' if we apply Barad's agential realism for quantitative psychological science.

## 2.4 Full causality and determinism

The understanding of causality (Table 1) is that a vector of influence is transported from one process, state, or object to another process, state, or object. A cause generates an effect on another entity or process. The philosophy of science perspective of quantitative psychology is built on this understanding. It is assumed that causal processes happen in the world that researchers investigate, and causal processes are used to discover real-world processes. So, specific determinants are assumed to transport vectors of influence to occurrences like "racial bias" and, for instance, generate or modify it (like 'empathy reduces racial bias'). For

2 Researchers, of course, have other responsibilities (like handling research participants well), but these are outside the scope here.

the discovery of real-world processes, the basic idea of an experiment is that different behavior or experiences between different conditions can be attributed to the differences between the conditions as their causes.

Determinism is the idea that all events are causally determined and that every outcome has at least one cause. As known, a deterministic system is characterized by the fact that previously existing causes unambiguously and completely determine its state in the future. Within determinism it holds that: If we find any variance empirically, there must be causes for this variance. If we cannot explain a variance, it is always attributed to epistemic reasons: we do not know enough about the determinants that cause the variance. Total determinism does entail that, ontically, there are causes for every outcome. Now, philosopher of psychology Gadenne states that strict and total determinism is not tenable for psychology; however, Gadenne argues that this is because of the *inexplicability* of chaotic processes and not due to indeterministic processes. The statement "chaotic processes follow strict causal laws, but are bounded by explicability and predictability" (Gadenne, 2004, p.125, my translation) exemplifies that Gadenne's reason to question determinism as tenable is only epistemic—i.e., not being able to *know* about all determining influences. Gadenne does not assume indeterminacy ontologically. I argue that the research logic of quantitative psychology is built on total determinism, and all variance is attributed to epistemic issues. Even the unsystematic variance of every measurement outcome is understood as part of a so-called measurement error. Likewise, the attempt to gain more and more objectivity resembles the understanding that, in principle, there is a cause for every variance and that we are 'bounded only by predictability' (see Gadenne above).

## 3 Agential realism as the philosophy of science perspective for quantitative psychology

This section introduces Barad's agential realism for the field of psychology. Barad was trained as a theoretical physicist and presents the alternative philosophy of science perspective with reference to physical objects, measurement processes, measurement outcomes, causal linkages, etc. This vocabulary makes the reasoning somewhat accessible for quantitative psychologists. Like them, Barad is talking about experiments. However, the agential realism perspective entails fundamentally different conceptualizations of science's objects, processes, and outcomes (see some comparisons in Table 1 and more in detail explained hereafter).

"Knowledge's are not innocent representations, but intra-actions of natures-cultures: knowledge is about meeting the universe halfway" (Barad, 1996, p. 189).

Barad negates the idea that with science, we find representations of real objects. Instead, Barad approaches realism concerning *entangled phenomena*. Before I enter the clarification of specific concepts, I look at the name 'agential realism.' Barad chose the term 'realism' because their aim is still to approach the 'nature of nature' or 'nature of reality.' The target is explicitly not "a matter of human experience or human understandings of the world" (Barad, 2007, p. 160). Barad chose the term 'agential' because this 'nature of reality' is understood as co-constructed by agencies (which need not be human). The underlying reasoning should become apparent after describing Barad's framework and its possible application within quantitative psychology.

Barad draws heavily on the 'philosophy-physics' from Niels Bohr (although departing from it in specific issues). Barad examines at length the arguments between Bohr, Heisenberg, Einstein, and some of their colleagues in the 1920s and 1930s about some physical experiments. The arguments led to the famous Copenhagen interpretation of quantum phenomena, for which both Bohr and Heisenberg are held responsible but which is not in focus here. Despite their commonalities, there is a specific difference between Bohr's and Heisenberg's understanding of the 'nature of nature', which Barad draws upon. Both agree upon "the final failure of causality" (Heisenberg, 1927, p. 83), and they agree on this point: "what is wrong in the sharp formulation of the law of causality, 'When we know the present precisely, we can predict the future,' is not the conclusion but the assumption. Even in principle we cannot know the present in all detail" (Heisenberg, 1927, p. 83). The critical difference between the conceptions of Bohr and Heisenberg is the *reason why we cannot know* (the present) in all detail. Barad carves out that Heisenberg attributes the source that we cannot know in all detail to *epistemic reasons,* while Bohr attributes the source to *ontic reasons*. According to Barad, Heisenberg refers to a disturbance in the measurement process and centers on 'possibilities of measurement.' This disturbance in the measurement process is an epistemic question. Bohr, by contrast, centers on 'possibilities of definition' as an ontic question (see Barad, 2007, p. 301). Barad follows Bohr and assumes a certain *indeterminism at the fundamental ontic level of existence*. This is the first peculiarity of agential realism. Within this framework, the uncertainty 'not to know in all detail' is *not* due to epistemic problems but is indeterminacy at an ontic level.[3] Importantly, this indeterminacy can be resolved. After a resolution, there are determinate states in the present, but they are also contingent on the (experimental) configurations. This is the second peculiarity of agential realism. A bounded object does not exist *per se* but is an outcome of a process; larger configurations resolve the indeterminacy into a determinate state. As a principle, these larger configurations belong to the outcome. These points and some corollaries are explained further in the following sections.

### 3.1 No entity realism, but realism toward phenomena

Agential realism does not assume individual objects with determinate boundaries or properties with determinate meanings as pre-existing *but as outcomes of processes*. By definition, this is no entity realism. Instead, reality "is composed not of things-in-themselves or things-behind-phenomena but of things-in-phenomena" (Barad, 2007, p. 140). This is to assume ontologically a thoroughly relational existence of everything; nothing exists independently by itself: "there are no independent relata, only relata-within-relations" (Barad, 2007, p. 429,

3 This does not mean that agential realism is based on Bohr's 'complementarity' ("simultaneously necessary and mutually exclusive"; *cf.* Barad, 2007, p. 415); rather, complementarity (and all its consequences) also follows from this pre-assumption of fundamental *indeterminacy* instead of a *disturbance in the measurement process*.

footnote 14). In philosophy of science, the term 'relatum' (from Latin) refers to the object to which a relation proceeds; the plural form is 'relata.' In the examples 'I see a particle' and 'I see empathy', particle and empathy are each a relatum. Barad wants to express that there is no relatum without relations. To apply agential realism in psychology, I use 'entity', 'object', and 'relatum' interchangeably (see Table 1). The point here is, that none exists without relations. This builds on Bohr's insight that, on the ontic level, the 'nature of nature' exists only in relation to specifics of (experimental) configurations. The reason that we can still deal with present individual objects is an occurrence that Barad calls *intra-action* (Table 1). Barad chooses this neologism (in contrast to 'interaction', Table 1) to express that intra-acting relata are not separate entities that influence each other or that one entity influences the other. Relata, objects, and entities do not preexist relations. As pre-assumption about the world, we should not assume any object—physical particles in the same way as psychological research objects like empathy—as existing without relations. The whole—a relatum and its relations—emerges only through specific intra-actions. Therefore, intra-actions enact 'agential separability.' What we see as boundaries between two seemingly separated relata, objects, or entities are *agential cuts* (Table 1). Barad uses the adjective 'agential' to express that the correspondent referent—'realism', 'separability', or 'cut'—does not exist *per se* but that an agency brings this correspondent referent into being. The separability does not exist by itself but is agentially enacted. Intra-actions enact an agential cut but are themselves 'agential.' Importantly, there is no inevitable 'agent' behind the agential becoming. Humans can be agents but are not required. The term agential is just a marker for the understanding that a distinction between separate entities is an *effect* "in contrast to the more familiar Cartesian cut which takes this distinction for granted" (Barad, 2007, p. 140).

This fundamental relationality also applies to psychological research objects and entities such as Dienes (2008) examples 'thoughts, images, networks, social pressures, social identities.' Certain relations are crucial for the very existence of any relatum, and if these crucial relations are different, then the relatum is different. The relationality also applies when psychologists state that a concept like 'social pressure' is just like a molecule composed of much smaller atomic parts like 'self', 'others', 'social norm', 'observable behavior in relation to that norm', etc. That is, we should not think of any component as a distinct entity. Even if psychologists try to differentiate an occurrence like 'social pressure' into its assumed parts, according to Barad, no part exists without enacting relations. Relata and their relations are a conglomerate. That which is understood as an entity from the contemporary, canonical perspective is understood as an entity-within-phenomena or thing-within-phenomena (Table 1) from the agential realist perspective.

Phenomenon (Table 1) is Barad's term for such a conglomerate. It is the name for the conglomerated relations. "It is through specific agential intra-actions that the boundaries and properties of the components of phenomena become determinate and that particular concepts (…) become meaningful" (Barad, 2007, p. 139). That means two entities might interact—as situated (!) separate entities—but their separateness is due to the encompassing phenomenon. The other way around is a phenomenon "produced through complex agential intra-actions of multiple material-discursive practices" (Barad, 2007, p. 206).

Instead of presuming entity realism, agential realism takes *phenomena* as the primary ontic unit of reality. This perspective recognizes a 'reality' (Table 1) in the sense of a shared experience. That is, discrimination can be 'real' (Table 1) for specific people and not for others. A physical object can be 'real' for specific radiation and not for others. However, this shared experience is presumed to be always bound to situations. This way, agential realism assumes we can encounter a situated reality, but composed of 'real phenomena' rather than composed of 'real entities.' Necessarily, a situated reality is bound to time and place. The term phenomenon refers to the relationality of each occurrence. Phenomena include the specifics of (experimental) practices and all these relations that are part of what seems to be a 'real object' in situations. Phenomena encompass the entanglements that enact a relatum-within-relations. Every noun we use in psychology, every psychological research object—for example, 'thought', 'social pressure', 'sensibility', 'attitude', 'self', and 'stereotype'—must be understood as object-within-phenomena. To be more precise, we could call something an 'object-within-phenomena' when the contemporary perspective understands it as the smallest part and as the component of an occurrence. Likewise, we could call something a 'phenomenon' when the contemporary perspective understands it as an occurrence that is composed of smaller parts. However, I do not promote such a differentiation. After all, from the agential realist perspective, we do not need it because everything is entangled. For example, Gemignani et al. (2023) applied this agential realist perspective by understanding 'migrants,' 'feminists,' 'oppressed,' or 'social justice' as such phenomena.

## 3.2 Entanglement of ontic existence and epistemic approach

Understanding everything that we name with nouns as relata-within-relations—respectively as 'phenomena'—and acknowledging indeterminacy until intra-actions place agential cuts implies that epistemic approaches cannot only discover what is already out there. Barad builds on Bohr's insight that the configurations of an experimental apparatus (Table 1) co-create the outcome. These configurations are the compositions of material and discursive settings that enact certain intra-actions, not others. In physics, this can mean more, but not exclusively, material than discursive configurations. However, not surprisingly, there is no clear differentiation between material and discursive within agential realism. Barad explains an example of a cigar being necessary for the outcome of a specific physics experiment. However: "Not any old cigar will do: the high sulfur content of a cheap cigar is crucial. Class, nationalism, gender, and the politics of nationalism, among other variables, are all part of this apparatus" (Barad, 2007, p. 165). That is, the social category of gender is entangled with an 'object' like a cigar. Class is entangled with the necessary high sulfur content, and so on. The agential realist perspective sees these variables not as separate influencing forces but as parts of the apparatus. Accordingly, specific experimental apparatuses (Table 1) with their specific material-discursive configurations enact specific agential cuts. Likewise, epistemic approaches—in which researchers use specific apparatuses—enact intra-actions that set agential cuts.

A psychological apparatus (i.e., a psychological 'instrument'), such as a questionnaire, has its own material-discursive configurations. Configurations of the apparatus itself—for example, the wording of questions—and configurations that enabled this apparatus beforehand.

For example, we find historical and social changes in attitudes embedded in the logic of specific questions within questionnaires. Similarly, material phenomena, like the availability of telephones or the internet, at specific historical periods are embedded in the changes in attitudes and so on. This way, the apparatus itself is not a bounded entity. It is entangled in material-discursive configurations. Of course, every part of the apparatus is. There is no part, component, entity, or object of the apparatus not entangled with material-discursive configurations. Likewise, every part of an apparatus—like the wording of questions—is itself a specific configuration that enables specific agential cuts. When an agential cut is placed, this is the moment where the ontic indeterminacy is resolved into a situated reality. That means apparatuses are productive. In physics, they can materialize an object. In psychology, we might prefer to say 'realize an occurrence.' As said, this occurrence can be any research object or anything we name with words. Of course, this realization process is not restricted to experimental settings but is part of the ongoing dynamic reconfiguring of the world. For example, a specific questionnaire will—as apparatus—realize outcomes in an experiment, and a specific wording in cultural stereotypes will—as material-discursive configurations—realize occurrences in schools. Outside of (laboratory) experiments, Barad's term 'material-discursive configurations' might be more suitable than 'apparatuses'; for what they do, they can be used interchangeably. The important effect is that "apparatuses are not mere observing instruments but boundary-drawing practices" (Barad, 2007, p. 140). So are material-discursive configurations. If these configurations are co-creating what exists situatedly, then the ontic existence is not independent of the epistemic approach. Every epistemic approach establishes specific configurations and not others.

Agential realism assumes an inextricable entanglement of ontic existence and epistemological approaches. Then, by definition, a measurement outcome (Table 1) cannot be an 'innocent representation' of an independent truth (see section 3). If our apparatuses participate in realizing outcomes, then the measurement process (Table 1) is partly a creation. That is, to measure is not only an epistemic activity. To measure is an intra-action which leaves boundaries. To measure carves out one of the several possibilities. "The point is that measurement resolves the indeterminacy" (Barad, 2007, p. 280).[4] The larger configurations, the material-discursive practices 'take' a measurement and produce agential cuts. To try to achieve independence of preferably every influence, as quantitative psychologists mostly do, is a fundamentally different approach than to assume that nothing exists without relations. In the agential realist perspective, we can never distance ourselves from that with which we intra-act. Importantly, we can never distance ourselves—not for epistemic reasons alone but because ontic existence is entangled with the epistemic approaches to it. This entails consequences for criteria of quality for science, which will be addressed in section 5.

"Practices of knowing and being are not isolable; they are mutually implicated. We do not obtain knowledge by standing outside the world; we know because we are *of* the world. We are part of the world in its differential becoming" (Barad, 2007, p. 185). Because one cannot disentangle knowing (Table 1) from being or epistemology from ontology, Barad again uses a neologism to explicate this point: "Onto-epistem-ology—the study of practices of knowing in being—is probably a better way to think about the kind of understandings that we need to come to terms with how specific intra-actions matter" (Barad, 2007, p. 185). Barad encourages us to understand the processes of emergence and not resign in the face of this entanglement. However, we need to learn onto-epistemology instead of continuing to try to approach an object 'as neutrally as we can', because neutrality does not exist—onto-epistemically.

There is also the issue of researchers' placing agential cuts and, therefore, taking part in the world's differential becoming. This impact goes beyond the handling of research outcomes (like the handling of outcomes discussed in the realm of the atomic bomb) but operates at the level of co-creating the research outcome itself. This is why Barad also demands to imply ethics: "[W]hat we need is something like an ethico-onto-epistem-ology—an appreciation of the intertwining of ethics, knowing, and being" (Barad, 2007, p. 185). Notably, the concrete ethical line cannot strictly be derived out of agential realism, only that we should fundamentally imply ethical considerations because we cannot purport that we only discover what is out there but take part in the formation of what we 'find' or 'create', respectively. This implicates the responsibility that researchers have.

## 3.3 Responsibility of the researcher in agential realism

Within agential realism, researchers are not only responsible for the kind of knowledge that they seek "but, in part, for what exists" (Barad, 2007, p. 207). To the degree that human practices are involved in the intra-active becoming of the world, humans are agentive participants and co-creators of the world. Notably, within agential realism, this boundary-drawing is not an unwanted influence that must be eliminated but an inevitable part of the phenomenon. If researchers agentially set boundary-drawing apparatuses, they have to question what exactly they 'measure.' If measuring is carving out one of several possibilities, researchers chose one particular possibility. It became famous that Isaac Asimov (1920–1992) doubted if 'intelligence' is just that, what the 'intelligence test' measures. Within agential realism, psychologists must doubt if constructs are indeed 'just that, what the test measures.' We cannot conduct any psychological experiment, exploration, or analysis without using a particular line of thought, specific language, certain nouns and verbs, maybe pictures, graphs, or icons. All these, too, are embedded in their social-material-historical entanglements. They have specific meanings for certain people in certain constellations at certain times and places and other meanings in others. Accordingly, researchers' decisions about design and material, along with all their histories and entanglements, are also part of the boundary-drawing practices in research settings.

These decisions of researchers can have strong and far-reaching consequences. For instance, Teo (2008, 2010) refers to Spivak (1988)—who coined the term 'epistemic violence'—and transferred this as 'epistemological violence' to psychology to stress that this violence is executed in knowledge production. Teo concentrated on situations where interpretations of data "implicitly or explicitly construct the

4 There is also a discussion of whether there are situations where such a resolution can be undone (see Barad, 2007; Schrader, 2012), but these situations and discussions are beyond the scope of this text and could be explored elsewhere.

*other* as inferior or problematic, despite the fact that alternative interpretations, equally viable based on the data, are available" (Teo, 2010, p. 298, emphasis in the original). One could expand this logic and name it 'epistemological violence' whenever researchers' decisions within study designs create negative consequences for some people. It is this important shift in the idea of research that researchers are partly responsible for what exists. Then, researchers have to include ethical considerations, such as developing criteria for judging what a 'negative consequence for people' is, in order to derive rationales for decisions about research design and material.

At the same time, it is important to state that researchers do not have full control over an outcome: "not everything is possible at every moment" (Barad, 2007, p. 182). Researchers do not have the opportunity to do every intra-action they might want. They are themselves just a (relational) part of the material-discursive practices. They are researchers-in-relations and use language-in-relations and experimental designs-in-relations with groups-in-relations and so on. With all their agentive practices, researchers are neither fully responsible nor not responsible for co-creating the outcome. Does this allow us to ask how big the researchers' 'part' is? If there are either previous or non-researcher agential cuts, could we then not ask: where is the line between researchers' responsibility and their non-responsibility? Such a question might arise from the hope of being able to distinguish between situations where experimental design decisions have an impact on the shape of the outcome and situations where they have only too little or no impact. It is the idea that configurations other than the researchers' have set most of the agential cuts, and the researchers' possibilities for influence are negligible. Suppose there are, in principle, phenomena where researchers cannot sufficiently co-construct the outcome. In that case, it seems reasonable to try to distinguish such phenomena from others where the construct is just that, what the test measures (see beginning of section 3.3). However, the notion of an 'extent of influence' resembles the conventional idea of a possible separation between a phenomenon and the influencing configurations or between the 'humanly discursive' and the 'non-humanly material' practices. Instead, Barad states: "Indeed, it is through such [material-discursive] practices that the differential boundaries between humans and nonhumans, culture and nature, science and the social, are constituted" (Barad, 2007, p. 140). Any separation we find as an outcome is an "agential separability—an agentially enacted ontological separability within the phenomenon" (Barad, 2007, S. 175). This means the separation is not pre-existing and just there to be found, but different configurations will enable different possibilities for agential separation. This leaves us with a reasonable desire to disentangle different influences because we have the hope to decide where researchers can do 'better.' At the same time, we have to acknowledge that such a differentiation is onto-epistemical—and onto-epistemological (see Table 2)—impossible because of the inextricable entanglement of material-discursive practices. Ways of implementing this understanding in agential realist psychology are discussed in section 4.3.

## 3.4 Reworked causality and emerging possibilities

Incorporating indeterminacy and fundamental entanglement in a thinking model does not mean there is no causality. However, this means reworking the previous canonical understanding of causality (Table 1), which is about the 'relationship between distinct sequential events.' Agential realism rethinks this in terms of intra-activity: "Intra-actions do not simply transmit a vector of influence among separate events. It is through specific intra-actions that a causal structure is enacted. Intra-actions affect what's real and what's possible, as some things come to matter and others are excluded, as possibilities are opened up and others are foreclosed" (Barad, 2007, p. 393). This way, we are invited to think of causality as entangled with conditionality. Causes also exist, but not as a sole reason but only together with conditions that render a causal chain possible (and then any given outcome is not the only possible). There is a causal impact in intra-actions but not as the 'transmit of a vector of influence among separate events.' It is sort of a thinking of neither 'anything goes' nor 'total determinism' when Barad talks of the "open-ended becoming of the world which resists acausality as much as determinism" (Barad, 2007, p. 182). In Barad's ongoing becoming, we find both the renunciation of then-separate entities and the implementation of context, conditions, and configurations that enable a causal execution.

This perspective encompasses both the confinement of possibilities and the multiplication of possibilities. "Intra-actions reconfigure the possibilities for change. In fact, intra-actions not only reconfigure spacetimematter but reconfigure what is possible" (Barad, 2007, p. 182). On the one hand, possibilities for an outcome are confined by intra-actions that set certain agential cuts. Confronted with such an agential separability, researchers cannot realize any outcome they might wish. This acknowledges that variables-in-relations can impact other variables-in-relations, and sometimes we cannot escape some impact. Quantitative psychologists are used to the idea that they only observe the interaction of variables. However, compared to a deterministic understanding, agential realism assumes the existence of several possibilities for an outcome. These possibilities (a) have partly no further reason because there is some fundamental indeterminacy, and (b) have partly the intra-actions as reasons, which can possibly be realized differently. That means there are two sources for the multiplication of possibilities for an outcome. While determinism holds that there are always reasons for a specific outcome, agential realism opens the question of where the world can be realized differently. The different consequences of applying issues (a) and (b) in quantitative psychology are enfolded in sections 4.1 and 4.2.

Agential realism also reworks the understanding of what is 'objective' (Table 1): "[O]bjectivity in an agential realist sense requires a full accounting of the larger material arrangement (i.e., the full set of practices) that is part of the phenomenon investigated or produced. (To do otherwise is to misidentify the objective referent). Hence objectivity requires an accounting of the constitutive practices in the fullness of their materialities, including the enactment of boundaries and exclusions, the production of phenomena in their sedimenting historiality, and the ongoing reconfiguring of the space of possibilities for future enactments" (Barad, 2007, p. 390–391). Objectivity, then, is about communicating the larger configurations of a boundary-drawing apparatus. Put simply, if we manage to inform colleagues about most of the involved relations of the investigated relata-within-relations, then we increase the possibility that they can reproduce these involved material-discursive practices and the realization potential of this phenomenon. Agential realist objectivity is not about eliminating influences but about communicating material-discursive practices as much as possible.

# 4 Consequences of applying agential realism in quantitative psychology

Barad proposes a meta-theoretical perspective, which I discuss as a philosophy of science perspective suitable for quantitative psychology. Many issues highlighted through the confrontation with agential realism have already been discussed (sometimes extensively). For example, to question the inherent boundaries of research objects and the necessity to take producing configurations into account has already been (and for a long time) discussed by Gergen and Gergen (1988, 2003), who propose that the self and every psychological construct are relationships rather than individual entities. K. Gergen implements this understanding in everyday life and professional practices, such as education, therapy, and knowledge production (Gergen, 2009). So, I propose agential realism not because it would offer completely new conceptualizations. However, it offers a set of bundled assumptions and corollaries that can be pretty accessible for researchers trained in quantitative logic because both approaches discuss experiments, objectivity, apparatuses, measurement, or knowledge acquisition (see also Table 1).

The pre-assumptions described in section 3 have several consequences for quantitative psychology, but here I focus on four: 1. The agential realist perspective changes the conception of a 'true score.' 2. It changes the conception of the context. 3. It changes the conception of the researchers' responsibility. 4. It changes the conception of the research endeavor.

## 4.1 Indeterminacy means there is no true score

The first issue arising from Barad's reasoning is the idea that there is a core indeterminacy in our world. This, however, is not an uncertainty due to the epistemic reason of 'not knowing well enough' but due to the ontic reason of 'not being determinate' at a certain level of existence. We can imagine that—in *certain situations*—this indeterminacy can be too small to have a relevant impact. However, which situations are concerned can be treated as an empirical question and cannot be a pre-assumptions. Further, the question of whether an impact is 'relevant' will again depend on the context and aim of the research. The agential realist pre-assumption is that a certain indeterminacy is part of the phenomenon a researcher is interested in. This indeterminacy is the reason for an unexplained variance.

Let us imagine repeated measurements under consistent conditions, at least in theory (because they are hardly realized in psychology). If we repeat any measurement, we will achieve varying values even if we do not change conditions. These variations are termed variance; alternatively, we utilize the square root of the variance, known as the 'standard deviation.' Traditionally, this variance is perceived as comprising both 'systematic' and 'unsystematic error.' The first is perceived as stemming from an unwanted influence, which has to be eliminated to approach the 'true' unbiased score. For the 'unsystematic variance', it is acknowledged that it cannot be eliminated, but it is conceived as indicating where the assumed 'true score' may lie. In this way, the traditional conception treats the unsystematic variance as stemming from pre-assumptions problems like 'not knowing well enough' where the 'true score' may lie. Consequently, the distribution curve of this variance is then used to infer the assumed 'true score' while admitting a little 'uncertainty.' However, from the agential realist perspective, this unsystematic variance stems from ontic indeterminacy. Consequently, there is no assumed single true score behind a blurred measurement process but a variance of possibilities. We can think of the distribution curve as showing the probability of each outcome but with the alteration of assuming an indeterminacy within a certain range instead of uncertainty about one true score. With an agential realist perspective, we have to assume a realization potential—i.e., the unsystematic variance distribution curve—for each specific configuration setting instead of one true score.

### 4.1.1 Every system has a realization potential

Whenever we psychologically 'test a person', we have to assume that there is an inherent variance within the investigated 'feature', which is actually a 'feature-system.' The term 'system' is added here to indicate that agential realism does not assume features that can be measured but that a feature is carved out of a larger phenomenon through material-discursive practices (including measuring devices). That we have to assume an inherent variance applies even if we could repeat the larger configurations of the situation exactly. This variance is not due to a measurement error but due to an inherent part of the whole phenomenon. When we obtain a particular realization, the variance is only broken down to a particular outcome value because of intra-actions, which cause this realization out of the larger potential of realizations. If we were able to repeat the same intra-actions, we would nevertheless obtain a more or less different result. Within the realization potential, this more or less different result has no further reason to be different but is only more or less likely. Hence, 'measuring' a 'feature-system of a single person' means obtaining information about this specific system's realization potential within the given configurations. Hence, researchers no longer search for an assumed single true score but for a range of realization possibilities.

### 4.1.2 Consequences for comparisons

Importantly, this conceptualization changes the comparison between people. Researchers then do not compare two different scores—no matter whether true or estimated—but we compare two distribution curves. Whenever these curves overlap, this brings similarities instead of differences to the fore. When comparing such potentials between persons, it is quite possible that in specific configurations and concerning a specific scale, person A shows a realization potential different from person B: different concerning the mean and/or different concerning the standard deviation and/or kurtosis of the realization potential. Whereas in the conventional understanding, the curves were used to deduce a significant difference, this agential realist conceptualization highlights the overlap of the two distribution curves. Within the overlap a difference as well as no difference can *occur with no further reason* other than the ontical indeterminate variance. If the realization potentials of two feature-systems overlap, then in agential realism, researchers do not consider this as an 'inner difference' but as 'a sometimes realized difference and a sometimes realized sameness.'

### 4.1.3 Consequences for replication from indeterminacy

The first critical consequence for replicability is that, from an agential realist perspective, the replication rate has an onto-epistemic

limit. A replication rate of 100% is, in principle, not possible for onto-epistemic reasons. It is not that theoretically 100% was the ideal that we cannot reach for epistemic or practical reasons, but that incorporating a fundamental ontical indeterminacy limits the possible replication rate for each phenomenon. The rethinking of 'true scores' in the form of realization potentials has a critical impact. Realization potentials increase the overall variability of results for any setting of conditions. Suppose we have to aggregate data of a group of people. In that case, we do not assume the property as having a determinate value somehow 'inside the person.' Instead, we assume kinds of 'individual phenomena'—with all their entanglements—which we aggregate. The idea is, in some situations, it might be possible to realize a specific property, but in principle, this happens in the form of a specific variance. Let us continue to theoretically assume that we were able to repeat the same configurations. If we incorporate the indeterminate variance of each 'individual phenomenon' and analyze a group of people, this understanding increases the ontically caused variance of the group. This is because every individual brings more than only one value to the group's variances.

Concerning the replication of studies, I argue that until now, psychologists have taken an outcome as indicating an assumed true score instead of one of the possibilities. This is relevant when comparing two outcomes, for instance, from an older study and a replication study. A replication of psychological studies often tries to replicate a significant difference between two groups. The replication fails if the significant difference was shown in a previous study but not in a recent study. Suppose we assume realization potentials instead of single true scores. In that case, it might well be possible that both outcomes, the difference outcome and the no difference outcome, are part of their regular group-comparison realization variance. Imagine we could estimate each group's whole regular realization potential, and both curves might overlap partly. If we realize an outcome of each group in one study, in most cases—unless the distributions hardly overlap, but even in cases where the distributions totally overlap—this applies: Finding a significant difference has one particular possibility, and finding no significant difference has another particular possibility. Importantly, in the agential realist perspective, this stems from ontic reasons and not from epistemic ones. To find a difference or not with specific possibilities belongs to the phenomenon and is not a measurement error. We are neither supposed to find a significant difference in each comparison nor to find no difference. The realization of a difference and the realization of no difference can be part of the configurated possibilities. Accordingly, if we (at least theoretically) replicate the same configurations many times, we could look at a proportion of the realization of differences and a proportion of the realization of sameness. By that, we judge a replication study's outcome differently than up to the present. It is mostly no longer about whether mechanism A exists or not. It then is about the question of how often—out of the ontic possibilities—mechanism A might realize (within these configurations) or not.

Of course, other problems related to a low replication rate—like publication bias, flawed research, or false positives—still exist. However, from the agential realist perspective, another issue is added: A low replication rate of a specific mechanism does not necessarily prove that phenomenon P does not exist, but it can prove the regular indeterminate variance of this phenomenon, that, for example, sometimes results in a group difference and sometimes not. Researchers have to discuss the basic idea of replications and the meaning of replication study outcomes anew when applying the agential realist perspective.

## 4.2 Configurations as part of things-in-phenomena

A second important alteration in thinking arises from Barad's reasoning that the relations are always already part of the relata. These ralata-within-relations only exist due to configurations (or larger apparatuses) that set agential cuts via intra-actions. Agential realism assumes that the objects and outcomes we find are materializations/realizations of material-discursive practices. Outcomes are sedimentation of intra-actions of these practices, which cause agential cuts. This framework states there is no such thing as 'ingroup-favoritism' or a 'representation of the other' in our mind without larger configurations co-creating it and indeed being part of what we named 'favoritism', 'representation', or any other construct. There is no thinking, feeling, or behavior without material-discursive practices that set agential cuts around a then-named thought, feeling, or behavior. This is a fundamental shift toward a relational ontology. Not a single psychic phenomenon exists without its history of entanglements and ongoing material-discursive intra-actions enacting agential cuts. Nevertheless, we should not misunderstand Barad's phenomena as deterministic systems and should not repeat the search for determined causal chains within them. As mentioned, even causality is something enacted through specific intra-actions. With that in mind, we cannot treat context as a third influencing variable.

### 4.2.1 Context is not a third variable

The agential realist perspective also implies an alteration of psychologists' understanding of their objects 'in context': "The notion that human psychology [psyche][5] is shaped by the social context has been the central premise of the field for nearly a century" (Van Bavel et al., 2016b, p. E4935). However, the idea that the context 'transmits a vector of influence' toward human cognition fundamentally differs from Barad's notion of intra-actively enacted agential cuts and the ongoing becoming of the world. Within agential realism, cognition is *not a relatum that is influenced by* its separate-from-the-object surrounding context. Rather, any cognition, feeling, or experience is a material-discursive phenomenon contingent on historical and actual configurations. Also, what psychologists understand as basic, universally human, not-social, or enduring is a contingent outcome of larger configurations. 'To be shaped by' is exactly *not* what Barad understands of relata-within-relations. Van Bavel and colleagues assume a classic causal influence, whereas Barad assumes a co-creation. Within agential realism, we can understand the human psyche and all its contents as entangled in material-discursive practices, and the objects of psychological (!) interest, as well as the psychic states, are realized differently within different practices.

Concerning the demand to take the context as part of the phenomenon into account, there seems to be a growing willingness, for instance, within social psychology, to attach more importance to surrounding configurations. Both within the replication debate (e.g., van Bavel et al., 2016a) and as a principle (Weber et al., 2023),

5 As mentioned above, to be more precise, these authors should have used 'psyche' instead of 'psychology' because, from the quoted sentence alone, we cannot be sure what exactly is meant here, but from their text, we do know they are actually talking about the psyche.

psychologists have promoted that we need to understand psychology's objects as context-sensitive and context-embedded, respectively. Pettigrew even suggests celebrating: "Contextual social psychology is finally emerging" (Pettigrew, 2018, p. 969). Jost promotes a new journal that embraces the context-embeddedness of psychical[6] phenomena because "we cannot stand outside of history—or culture or politics or economics" (Jost, 2024, p. 7). At first glance, this sounds like Barad's request. However, I want to stress two important issues: *First*, it makes a difference whether we conceive of this influence as a moderator variable, like a 'third' variable that 'transmits a vector of influence' and determines how the effect of variable A to variable B unfolds, or as inextricably entangled part of the phenomenon. Cultural psychology has extensively discussed their latter perspective as different from the conception of 'moderating.' Cultural psychology, as, for example, Chakkarath (2011) and Chakkarath and Straub (2020) describe it, is based on the principle that culture and psyche evolve through a reciprocal, mutual co-construction. Psychic structures, processes, and functions are understood as inherently entangled with cultural lifestyles, practices, languages, and discourses and non-existent without their context.[7] *Second*, and this point might be a consequence of the first one, psychologists often apply the embeddedness primarily to the cognition or emotions of the research participants and rarely to the researchers. For instance, Pettigrew does not transfer the insight that "cultures and social norms moderate basic psychological processes"[8] (Pettigrew, 2018, p. 963) to the idea that researchers, their apparatuses, and the language and concepts they use are also always entangled within cultures and social norms—and what this entanglement means for the research process. Weber et al. (2023) acknowledge that "researchers are themselves embedded in systems of knowledge production," but that, importantly, is their final sentence and not the basis of their reasoning.[9] The most far-reaching application of the idea of embeddedness within social psychology seems to be undertaken by Cikara et al. (2022). They discuss various possible contexts/configurations, including "political, legal, research and regulatory institutions" (p. 545) as productive for social categories. They explicitly include the researcher's responsibility and address "the authoritative power given to science to shape truth and knowledge" (Cikara et al., 2022, p. 537). I reckon that their recommendations about study design and analysis choices can be founded on agential realism. Even though some approaches to context-sensitivity do not go as far as agential realism, a future agential realist psychology can still learn from such perspectives, for instance, regarding the application of specific methods. Pettigrew (2018) and Jost (2024) advocate multilevel modeling to link different levels of complexity. Skinner-Dorkenoo et al. (2023) demonstrate a systemic approach to racism. A thorough examination of such methods in relation to the basic assumptions of agential realism is one of the future tasks for agential realist psychology, which I address in section 5.

### 4.2.2 Consequences for replication from entanglements

As a consequence for replications, we must always consider the configurations of an outcome as part of our research questions and objects. We shift the focus to understanding an outcome beyond previously assumed inherent features (such as a person's characteristics) to encompass instead the entire producing phenomenon, including configurations previously considered outside of the investigated feature. This includes historical, material, and researchers' entanglements. To replicate 'ingroup-favoritism', for instance, we must consider the larger relations that render such an outcome possible. This often may require replicating those relations as well. Barad states: "Crucially, the objective referent of measured values is phenomena [*sic*], not (some abstract notion of) objects (which do have an independent existence)" (Barad, 2007, p. 340). We remember that psychological 'objects' need not be physical ones but can also be characteristics, etc. Accordingly, we must try to replicate situated phenomena using the reasoning described in section 4.1 rather than trying to replicate essential mechanisms or characteristics that are assumed to be universal.

For different reasons, psychologists have already discussed how neglecting the context can reduce the replication rate (e.g., van Bavel et al., 2016a), although discussions are about objections to context relevance and contingency circumstances. For instance, Landy et al. (2020) demonstrated the importance of operationalization choices for obtaining the same or at least similar results after conducting a study. Nosek et al. (2022) give a sophisticated overview, but they part from agential realism in important points. For one, they assume a total determinism, which can be read out of reasoning like "[An outcome reproducibility failure] can occur because of an error in either the original or the reproduction study" (Nosek et al., 2022, p. 721). This reasoning contradicts the inclusion of a fundamental indeterminacy explicated in section 4.1. Besides the mentioned onto-epistemical limit for a replication rate given existing indeterminacy, there are two more critical issues concerning the conceptualization of the context. The first is the necessity of considering the larger configurations. Nosek et al. (2022) promote caution against 'unconsidered factors' (p. 727), but they do not seem to see this necessity for every replication procedure. From the agential realist perspective, every finding is necessarily contingent on its configurations. Second, any 'unconsidered factors' and enabling conditions must be understood as outcomes-with-enabling-configurations in themselves. These factors are not variables with inherent characteristics or independent working processes that influence the primary object of interest. A 'racial bias' should not be understood as a feature of a person but as a culturally enacted phenomenon. It is a possibility within a culture system that is enabled through configurations. This cultural possibility has many more components-in-relations that must be accounted for. This accounting should not proceed deterministically, assuming bounded entities that have characteristics. It is not that 'racial bias' is a feature of the culture either. Features are not to be located within an 'object', no matter if the object is a person, a family, or a culture.

6 Originally named "context-embeddedness of social psychological phenomena" (Jost, 2024, p. 5).

7 This position is posed in contrast to cross-cultural psychology, which understands context as conventionally influencing the inherent processes of the human psyche. How the cultural psychology position is applied to concrete concepts is, for instance, demonstrated by Glaveanu (2014), addressing creativity, and Salter et al. (2018), addressing racism.

8 Significantly, Pettigrew actually means 'psychical processes' and not 'psychological processes'.

9 Similarly, Greenwald (2012) puts the insight that researchers are influenced in the final sentence instead of starting from that, though the text is titled "Scientists Are Human." The strategy, then, is not to overthink research processes but to try even harder to overcome such influences.

From an agential realist perspective, too much information is lost when researchers do not account for larger enacting configurations of a phenomenon. Furthermore, important information gets lost if researchers search for essential characteristics or 'vectors transmitting an influence' instead of co-creation. If the investigation is directed toward transmitting vectors, it misidentifies the investigated referent. As the first step, instead of eliminating 'influences', researchers must work *with* them. As the second step, researchers also need to search for enabling configurations in a nonessential way. This influences the idea of the research endeavor and touches on the question of generalizability. We can no longer think of realizations as widely valid as classic approaches assume, which presuppose that realizations exist independently and are merely biased by context. If a finding is a co-creation of relata-within-relations, then generalization is in question. I address this in section 4.4.

## 4.3 Tasks for responsible and accountable researchers

Section 3.3 clarified that researchers have a broader responsibility with an agential realist perspective than a classic approach. A new responsibility is added to previous responsibilities (like honest behavior, transparency, etc.) because researchers' decisions may play a part in the phenomenon's becoming. In this section, I discuss where to put some attention when we implement this understanding of the possible ontic involvement of researchers. If there is no underlying separation between the researcher's influence and non-researcher configurations that can potentially be detected, then we will rarely try to find such a demarcation line and instead start to learn to deal with this entanglement. There can be co-creations from researchers' decisions that cannot be eliminated because they are an inherent part of the phenomenon. That is why researchers cannot only rely on the strategy of trying to eliminate their part-taking. Part-taking must be fundamentally acknowledged and concerned with the following (amongst others).

### 4.3.1 Decisions must be justified

One consequence of this alteration is that findings are not as widely valid as classic approaches assume, which presuppose that a characteristic of a research object is, in principle, independent of the researchers. In section 4.2, I already discussed the limitation of general validity due to the context relevance of each phenomenon. The outcomes' dependence on researchers' decisions is a further limitation. When researchers cannot declare that they only study what is already out there, then they have to declare why they are studying the phenomena in the way they do. Then, the question of how to design research is not only about operationalizing a research question in the best way to represent an assumed pre-existing characteristic but also about why researchers build and frame the parts as they do. Why do researchers use certain language, conceptualize something one and not another way, frame a question this way and not another way, etc.? When researchers acknowledge that they play a role in the research outcome, *every decision about a research design must be accounted for rather than being self-evident*.

### 4.3.2 Ethics must be made explicit

Researchers need new criteria for accountability. Classic criteria for research quality do not suffice here because, within the perspective of entity realism, there is no need to justify the framing of a question beyond the examination of whether a design is an appropriate method to represent what is already there. Within agential realism by contrast, there is the possibility that a phenomenon could be realized differently, and researchers have to justify why they take part in a particular becoming and not in another one. This again makes clear why Barad proposes that we need an ethico-onto-epistemology. However, it is already clear that agential realism does not prescribe *which* ethical lines should be followed. Researchers have to explicate their ethics (and maybe a scientific community starts to discuss agreements about ethical lines in specific times and places).

From an agential realist perspective, we must start with situated guidelines instead of generalized ones. For the scope of this paper, we might orient toward rights like the right to life, liberty and security of the person, and freedom of thought, opinion, and expression. Such guidelines will imply striving to eliminate violent or discriminatory research outcomes. If researchers agree on such rights and an outcome still diminishes the freedom of expression of persons, then researchers are co-accountable.

Categorization into groups may be a prominent example of psychology's part-taking in outcomes. From the conventional perspective, some research objects or persons supposedly possess common features that other objects/persons do not have (or to a significantly lesser extent). This is a common reason to categorize them into groups.[10] Applying the agential realist perspective, we do not locate features within distinct objects, so this rationale for categorization is not applicable. A category does not present itself as self-evident. Instead, we always have to explain the rationale for grouping people in a certain way. This does not make categories useless; we can have good reasons for categorizations, but we have to tell those. However, it stresses the relativity of categorization and demonstrates the contextuality. Again, this clarifies that we need ethical explanations for categorizations and cannot disguise that there are choices behind our groupings. We have to confront researchers with the danger of executing epistemological violence (i.e., violence executed in knowledge production, see section 3.3) because researchers cannot return to the statement that they have just found what is there, independent from them. The same applies to the identification of differences. Differentiation could be a meaningful and appropriate action, but this, too, is contingent and a realization-within-relations. Again, it demonstrates the need for ethico-onto-epistemological considerations.

## 4.4 Altered research endeavor

Initially, I described psychology's research (Table 1) endeavor as an attempt to describe, to understand, to explain, and, in some cases, to be able to change human thought, feeling, and behavior. This conventional research endeavor is about knowing (Table 1) the mechanisms that determine results. This way, researchers suppose that they can explain why things happened in the past and hope they can predict what will happen in the future. Kim (1999), in order to develop an alternative, described the contemporary, canonical research endeavor as an attempt to find the 'periodic table of basic human

10 Other reasons for categorization include common fate or similar fit to a requirement.

behavior.' The identified basic 'elements' could then be used to explain even complex human behavior. Additionally, researchers hope that knowing the mechanisms allows them to intervene and sometimes control an outcome, at least a little bit (e.g., to help somebody feel better). This reasoning is based on the deterministic idea that the system's state at one point determines the state later. The agential realist perspective alters this reasoning in two critical ways: It includes an indeterminacy within causal processes. Moreover, it understands the components of any system as contingent from its enacting configurations (plus the indeterminacy also within this enacting processes), so that we must assume a connection of everything with others (i.e., relata-within-relations). No entity or process is disconnected, and no system within the universe is enclosed and separate from the rest.

These points change the research endeavor. The indeterminacy within processes diminishes the predictability on an ontic level. This understanding establishes variance as a regular part of every mechanism. Again, in the agential realist perspective, not everything is possible, but in each situation and configuration, more than one outcome is possible—for onto-epistemic reasons and not as epistemic fallacy. Furthermore, other configurations might disable specific realizations and enable new ones. This implies that we search for possibilities instead of the one true result. If researchers find one realization, a question arises about what other realizations might look like. Then, research is not only about 'how it is' but also about 'how else can it be?' The agential realist psychology accounts for possibilities. It disengages the idea of finding human mechanics that will repeatedly work the same way. Instead, psychological research (Table 1) is about psychic and behavioral possibilities. Agential realist psychological research strives to describe, understand, and explain the *possibilities* of human thought, feeling, and behavior—within specific configurations (including those of the researchers). That entails that research can look at specific realizations, the configurations of these realizations (including researchers' configurations), and other possible realizations (and their configurations). This is an alteration of research questions.

### 4.4.1 Altered research questions

Agential realism alters research questions. I will consider three types of research questions in the following three paragraphs. *First*, the formerly common question about 'the character of X' can still be pursued. However, any outcome is an answer about a *local and temporary phenomenon*, and extra attention needs to be given to ask for the scope of this contingent realization. *Second*, agential realism shifts our attention to the character of the enacting configurations and does not locate 'the character of X' only within a bounded entity. For each situated realization, researchers are simultaneously provoked to ask: What enabled this outcome? *Third*, agential realism directs researchers' attention to what was disabled before and what else can be enacted. If relations render some relata possible and others not, we can investigate which other relata can be realized. Above all, researchers have to justify why they follow a specific research question, use a particular study design, and put a particular configuration of their research apparatus—nothing can be just a matter of course.

Concerning the *first* type of research questions, which is about investigating a local and temporary feature, we can note: "The line between subject and object is not fixed, but once a cut is made (i.e., a particular practice is being enacted), the identification is not arbitrary but in fact materially specified and determinant for a given practice" (Barad, 2007, p. 155). A relatum can become situated 'real', even though 'being real' is then not about being existent without an onlooker/interaction (see classic realism) but about situatedly shared experiences of intra-actions and cuts (see Table 1). Researchers can be interested in investigating a *situated property*, a *local quantity*, or a *temporary character* of an entity-within-relations. Especially so-called 'applied research' is used to deal with phenomena that might have an important, situational impact but are limited by their scope. In the same way, so-called basic research must develop an understanding of any investigation as research about realizations within *local and temporary* conditions.[11] Such (onto-epistemological) knowledge can be very interesting for certain people and specific goals. However, a psychological study cannot reveal something about every human.

Concerning the *second* type of research questions, which is to investigate the enacting configurations, in some areas, new research questions might emerge. Instead of concentrating on the realizations that we find in our worlds, we can and should also ask what creates them and what brings them into the world. Especially if we want to use an outcome, for instance, 'persons X react to Y with Z', then we need to know more about the enacting material-discursive practices since we cannot assume the mechanism resides within people. The characteristics of anything are more sensibly located in relations than in entities. Researchers can no longer search for essences because they are not located in an entity but instead are an outcome of enabling configurations that researchers can investigate.

Concerning the *third* type of research questions, investigating what else can be enacted is whether other realizations can be carved out of the possibilities. This makes realizations less self-evident. It opens up the question of whether things could be otherwise and if realizations could be different. If the boundary-drawing apparatuses have specific configurations, we can research if and what other configurations can enable. This understanding also can generate whole new research questions. For every finding, we could start to ask, 'Can it be different?' This links to Barad's reminder that ethics play a role in the researcher's decisions because if 'it can be different', then we need to answer the question, 'Which difference is desirable and why?' Besides the new perspective on changed configurations, this alters the understanding of any first outcome as not a given but as *one* possible situation.

### 4.4.2 Altered interpretations

The ethico-onto-epistemology of the agential realist perspective alters interpretations of outcomes. The alterations are implicitly mentioned in the discussion of altered research questions in the section above: Any outcome is interpreted as a local and temporary realization, and it is a different research question of how far it might spread. Any outcome is not interpreted as residing within a person or an entity but in material-discursive practices and configurations, enabling this outcome. Any outcome is interpreted as one possibility amongst others, and it is a different research question of how frequently different realizations emerge.

11 I suppose that this reasoning erases the distinction between 'basic' and 'applied' research, but that discussion is beyond the scope of this article and must be held elsewhere.

Transferred to social situations, this touches on numerous interpretations. For instance, that any outcome is interpreted as one possibility amongst others transfers an insight from 'the human psyche saves energy through categorization' to ‚the system-of-human-psyche has the capability to save energy through categorization.' This changes the point of energy-saving from a 'must' to a possibility. Such an outcome—to save energy through categorization—is one possibility amongst others, depending on the configurations of the material-discursive practices plus an indeterminate variance (until intra-actions carve out one particular situated outcome). It changes our view on 'psychic mechanisms' when we no longer search for the hard-wired program in brains and minds but see possibilities within configurations. Then, we can ask for the situations and configurations when people do not categorize to save energy. This perspective opens up for the change of statements like 'humans automatically perceive skin color and gender' to ask 'under which configurations do human-systems not perceive skin color and gender?' Every given realization is not the only possible one.

Furthermore, looking for enabling configurations can change the interpretation of a locus of control. A 'racial bias' is then located not only within a specific person but also in the structures of society, the current language, thinking models, narratives, etc. A score in an implicit association test for racial bias is then interpreted as an indicator of cultural associations and not only individual ones.

When we see realizations as local and temporary, we cannot interpret them as elements of an assumed 'periodic table of basic human behavior' (as criticized by Kim, 1999). This changes the idea of generalization. From an agential realist perspective, generalizability is not a goal *per se*. Instead, we have to assume there are constellations of material-discursive practices that spread across every human on earth and constellations with a much smaller scope. It is an empirical question of which constellation realizes where and how often. Needing to breathe oxygen with lungs might be such an earth-wide (nowhere near 'universal') configuration for humans; needing to reduce cognitive dissonance (*cf.* Festinger, 1957) might not be earth-wide. Notably, the outcome that the need to reduce cognitive dissonance is *possibly* an earth-wide human phenomenon could be an empirical finding. However, I suppose these earth-wide configurations are rare for the psyche and psychology. Instead, with an agential realist perspective, we do not seek generalizability but understanding a specific situation, including its indetermined realization potential. Landy et al. (2020) organized 15 research teams to test the same research question, each with its own operationalization. They showed overall 'how design choices shape research results' to learn how to approach generalizability. However, with an agential realist perspective, a project like that would try to use the divergence of the results to learn something about the specificities of each operationalization. Not Generalizability is the goal *per se*, but knowledge about local and temporary phenomena.

# 5 Further directions

To take agential realism as the philosophy of science perspective for quantitative psychology changes assumptions about ethico-onto-epistemological basics, changes the procedure of science, and interpretations of outcomes. Nevertheless, agential realist psychology does not turn away from quantitative research but instead aspires to change former Newtonian realizations of quantitative psychology. However, first applications of agential realism into quantitative research—this paper included—can only begin initial discussions. There is still work to be done to develop a thorough understanding of the alterations of concepts, reworking of methods, and reinterpretation of findings. This work includes a further rethinking of important concepts of research that could be considered only insufficiently or not at all here. It also includes concrete tasks like revising existing methods.

## 5.1 Concrete tasks at hand

If we further elaborate on agential realist psychology, we need to aptly develop language. For European-influenced countries, Gergen and Gergen (2003) already asserted that too few good terms can describe relational thinking. This situation itself is an agential cut-enacting configuration and has its impact. Nouns imply an essence that determines why something is called what it is called. As one strategy, in English, it is sometimes possible to make a verb of a noun to indicate the enacting instead of stating a being. 'To gender' a person transports another meaning than 'the gender' of a person. Another strategy—one that Barad used frequently—is using hyphens to link words and concepts together. Like relata-within-relations emphasizes the becoming of relata through their relations, linguistic constructions like 'feature-system' could indicate an understanding of entanglements. Nevertheless, changing language requires agreement between more people who use a language.

Another task will be to examine previous methods. The alteration of concepts makes it necessary to examine existing quantitative methods and their suitability for agential realist conceptualizations. I suppose the knowledge of methods that can provide information about entangled configurations is growing. However, such methods are still primarily implemented to try to delete 'unwanted influences' instead of working with entanglements. For instance, Hanfstingl (2022) discusses the combination of 'specification curves' with 'combinatorial meta-analyses' to gain information about the effects of researchers' decisions. Another example is the already mentioned project of Landy et al. (2020). It would be interesting to apply such methods to work *with* the entanglements as configurations that are part of the phenomenon rather than to apply such methods to be able to delete the entanglements as a disturbance from the overall picture. Furthermore, as mentioned, Pettigrew (2018) and Jost (2024) promote multilevel modeling to link different complexity levels. Agential realist psychology can learn from these methods, but it is necessary to examine them in relation to the basic assumptions of agential realism.

On the methodological side, there are already sophisticated recommendations to imply quantum probability theory (QPT) for the modeling of cognition, called quantum cognition (e.g., Pothos and Busemeyer, 2022; Busemeyer and Wang, 2015). Unlike agential realism, the quantum cognition perspective does not formulate an understanding of the ontic state of research objects or the consistency of our world but an understanding of the nature of human cognition. Quantum cognition offers a model for the working of human cognition; agential realism offers a model for the 'worlding' of our world. For instance, quantum cognition applies an ontic indeterminacy to decision-making processes but not to psychological research logic. Nevertheless, I am convinced that an agential realist psychology can learn enormously from handling probabilities within these approaches

because the researchers install QPT calculations due to the assumed indeterminacy and not because of uncertainty about where the 'true score' is. So, I encourage approaching these QPT calculations of quantum cognition, not because they are a good model for human cognition but a good model for the 'worlding' of our world. In contrast, although item response theory (IRT) relies on probabilities, it still follows the classic understanding of the existence of a latent trait (in an ontic sense), which has to be measured in ways as sophisticated as possible (epistemologically spoken). It does not assume a core indeterminacy as part of every outcome but uses probabilities for epistemic reasons of 'not knowing well enough' (see differentiation in section 3).

Furthermore, developing methods to gain knowledge about the *realization potential of situated configurations* seems necessary. In mechanics, researchers might be able to repeat the same measurement process many times, but in psychology, this is far more complicated. Researchers can just let a ball hit the detection screen repeatedly to get an idea of the distribution curve of these configurations. This kind of repetition obviously will not work with persons. We might want to differentiate the realization potential of the behavior of person-system A from that of person-system B while still incorporating their overlap. Currently, a measurement is taken as an indication of the 'true score' with a specific uncertainty, but can that measurement be taken as an indication of the realization potential? What else can help to gain information about which realizations are less likely for person-system A than are other realizations?

In addition, there is much more to say about replication from an agential realist perspective. If we reconceptualize findings as not telling something about a 'true score' but about material-discursive practices, then we must continue rethinking replication. How do researchers deal with the extra variance stemming from an ontic indeterminacy until intra-actions enact agential cuts? How do researchers interpret an outcome itself as part of a realization potential when it belongs to the larger phenomenon that realization A (e.g., a group difference) sometimes happens and sometimes does not?

Hopefully, psychologists will see many more tasks at hand to elaborate further on an agential realist quantitative psychology. This paper can only start some discussions; indeed, different discussants' backgrounds will enrich and differentiate the elaborations.

## 5.2 Further working out of conceptualizations

Other tasks are concerned with mapping out some already developed conceptualizations further. For instance, it became clear that agential realism demands the inclusion of ethical reasoning because researchers are also part of the material-discursive boundary drawing. It also became clear that for onto-epistemic reasons, we cannot distinguish between an influence from the researcher and no such influence, which deprives us of the opportunity to try to delete the former. We must learn to incorporate ethical reasoning and the researcher's standpoints transparently and constructively. We must work out forms of assembling perspectives instead of trying to find a perspective from nowhere. Because researchers are humans, this might lead to a new psychology of science that does not see the researcher's practices as erasable disturbances but as onto-epistemic entanglements.

Further elaboration is also required in the understanding of context as entangled relations and not as a third variable. The field of cultural psychology demonstrates how to execute this perspective not as a psychological sub-discipline but as a general perspective on phenomena (*cf.* Chakkarath and Straub, 2020). This shows some fundamental similarities to the agential realist perspective. For instance, taking embeddedness seriously means dropping essentialism concerning objects and categories. If we see the context as co-creating, we question experiments about social phenomena conducted at the computer. One task is to rework measurement designs with a fundamental inclusion of the context and the researcher's position as entangled parts. Such acknowledgments that researchers are also embedded must move from the end of papers—where Weber et al. (2023) and Greenwald (2012) put it (see section 4.2)—to the start of research. That is, research must be built upon the premise of entanglement.

Of course, all these alterations affect the quality criteria for research. Future tasks include elaborating on quality criteria for agential realist psychology. Objectivity has already been renewed by Barad (see section 3.4). Nevertheless, researchers can use clearer instructions about communicating material-discursive practices within both psychic and psychological phenomena. Moreover, the concept of reliability has to be revised, and the concept of validity. For example, Barad does not discuss the concept of validity, hardly uses the word, and if so, then in a conventional sense of indicating something with "limited" (Barad, 2003, p. 823) or "questionable validity" (Barad, 2012, p. 12). However, validity cannot be applied to measurement in the contemporary way of quantitative psychology to describe that a test measures 'what it aims to measure' and that a measurement process delivers a true (as possible) representation of an entity. When measuring is instead an intra-action that can resolve the indeterminacy into a determined state, there must be a non-representationalist form of validity. This new validity has to include the idea of a 'faithful account of a real world' (Haraway, 1988) but does not understand measurement as the practice of relating a number to a pre-existing quantity (see discussion of Trout's definition in section 2.2). Further discussions of an agential realist validity and reliability are needed.

This paper encourages psychologists to *reconsider what their knowledge represents*. With Barad's agential realism, a new proposal about 'intra-actions of natures-cultures' emerges: "Knowledges are not innocent representations" (Barad, 1996, p. 189). "Hence, (…) what is at issue is not knowledge *of* the world from above or outside, but *knowing as part of being*" (Barad, 2007, p. 341, emphasis in the original). Knowing (Table 1) is then not to have information about the state of something. Instead, "knowing is a matter of differential responsiveness (…) to what matters" (Barad, 2007, p. 380). If I 'know something', I can respond differently, but not because 'my bounded entity' can 'act independently' with 'having information.' Rather, I can respond differently because with 'knowing', I am part of possible intra-actions and part of material-discursive practices. Agential realism shifts 'knowing' away from cognition—which is another example of why we need adapted language for these understandings. It understands practices of knowing and being as mutually implicated. To 'know' is kind of 'taking part' and also to 'do.' Within psychological science, we must consider how our onto-epistemical and onto-epistemological practices intra-act and co-create realizations. Many psychologists want information in the first place to make the world a better place. With agential realism, we skip the idea of 'gaining information first' but proceed directly to try to realize better realizations—which, as we know

already, needs ethical lines to locally and temporarily define what is 'better.' Science (Table 1) will not detect 'deterministic causal structures' but will help to understand situated possibilities.

Of course, in this paper, I made agential cuts myself. Corresponding to the agential realist perspective, the aim is not to avoid those but to communicate them as well as possible. I hope this text is transparent about which line of thinking is followed at which point and where turns are taken so that colleagues can enter the reasoning and realize other or similar cuts from their perspective and entanglements. Moreover, I suppose some training is needed to consider the dimensions of agential realism. I suspect that most quantitative psychologists are trained in thinking models and language that support classic understandings. I propose we take some time to rethink and relearn, but I recommend to start now.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

JS: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Acknowledgments

The author thanks the reviewers and editor for their constructive and helpful feedback.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Barad, K. (1996). "Meeting the universe halfway: realism and social constructivism without contradiction" in Feminism, science, and the philosophy of science. eds. L. H. Nelson and J. Nelson (Dordrecht: Springer), 161–194.

Barad, K. (2003). Posthumanist performativity: toward an understanding of how matter comes to matter. *Signs J. Women Cult. Soc.* 28, 801–831. doi: 10.1086/345321

Barad, K. (2007). Meeting the universe Halfway: Quantum physics and the entanglement of matter and meaning. Durham: Duke University Press.

Barad, K. (2012). What is the measure of nothingness? Infinity, Virtuality, justice. Osfildern: Hatje Cantz.

Brown, D. (2020). Self-structure singularity: considerations for agential realism in critical psychology. *Soc. Personal. Psychol. Compass* 14, 1–11. doi: 10.1111/spc3.12569

Busemeyer, J. R., and Wang, Z. (2015). What is quantum cognition, and how is it applied to psychology? *Curr. Dir. Psychol. Sci.* 24, 163–169. doi: 10.1177/0963721414568663

Chakkarath, P. (2011). Psychologie und Kultur. Zur Problematik adäquater Fachverständnisse und adäquater Methoden. *Z. Kulturphilos.* 5, 327–342.

Chakkarath, P., and Straub, J. (2020). "Kulturpsychologie" in Handbuch qualitative Forschung in der Psychologie. eds. G. Mey and K. Mruck (Wiesbaden: Springer), 283–304.

Cikara, M., Martinez, J. E., and Lewis, N. A. (2022). Moving beyond social categories by incorporating context in social psychological theory. *Nat. Rev. Psychol.* 1, 537–549. doi: 10.1038/s44159-022-00079-3

Dienes, Z. (2008). Understanding psychology as a science: An introduction to scientific and statistical inference. Basingstoke: Palgrave Macmillan.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford, CA: Standford University Press.

Gadenne, V. (2004). Philosophie der Psychologie. Bern: Huber.

Gemignani, M., Pujol-Tarrés, J., and Gandarias-Goikoetxea, I. (2023). Social justice narrative research: from articulation to intra-action and ethico-onto-epistemology. *Qualit. Psychol.* 10, 552–564. doi: 10.1037/qup0000254

Gergen, K. J. (2001). Psychological science in a postmodern context. *Am. Psychol.* 56, 803–813. doi: 10.1037/0003-066X.56.10.803

Gergen, K. J. (2009). Relational being, beyond self and community. New York: Oxford University Press.

Gergen, K. J., and Gergen, M. M. (1988). "Narrative and the self as relationship" in Advances in experimental social psychology, Social psychological studies of the self: Perspectives and programs. ed. L. Berkowitz (Cambridge, MA: Academic Press), 17–56.

Gergen, M., and Gergen, K. J. (2003). Social construction: A reader. Thousand Oaks, CA: Sage Publications.

Gilbert, D. T., King, G., Pettigrew, S., Wilson, T. D. (2016). "Comment on "Estimating the reproducibility of psychological science". *Science*, 351, 1037. doi: 10.1126/science.aad7243

Glaveanu, V. P. (2014). Theorising context in psychology: the case of creativity. *Theory Psychol.* 24, 382–398. doi: 10.1177/0959354314529851

Greenwald, A. G. (2012). "Scientists are human: implicit cognition and researcher conflict of interest" in Psychology of science: Implicit and explicit processes. eds. R. W. Proctor and E. J. Capaldi (Oxford: Oxford University Press), 255–266.

Hanfstingl, B. (2022). Future objectivity requires perspective and forward combinatorial meta-analyses. *Front. Psychol.* 13:908311. doi: 10.3389/fpsyg.2022.908311

Hanfstingl, B., Uher, J., Edelsbrunner, P. A., Dettweiler, U., and Gnambs, T. (2023). Editorial: from "modern" to "postmodern" psychology: is there a way past? *Front. Psychol.* 14:1091721. doi: 10.3389/fpsyg.2023.1091721

Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Stud.* 14, 575–599. doi: 10.2307/3178066

Heisenberg, W. (1927). "The physical content of quantum kinematics and mechanics" in Quantum theory and measurement. eds. J. A. Wheeler and W. H. Zurek (Princeton: Princeton University Press), 62–84.

Herzog, W. (2012). Wissenschaftstheoretische Grundlagen der Psychologie. Wiesbaden: VS Verlag für Sozialwissenschaften.

Højgaard, L., and Søndergaard, D. M. (2011). Theorizing the complexities of discursive and material subjectivity: agential realism and poststructural analyses. *Theory Psychol.* 21, 338–354. doi: 10.1177/0959354309359965

Hollin, G., Forsyth, I., Giraud, E., and Potts, T. (2017). (dis)entangling Barad: materialisms and ethics. *Soc. Stud. Sci.* 47, 918–941. doi: 10.1177/0306312717728344

Jost, J. T. (2024). Grand challenge: social psychology without hubris. *Front. Soc. Psychol.* 1:1283272. doi: 10.3389/frsps.2023.1283272

Kim, U. (1999). After the "crisis" in social psychology: the development of the transactional model of science. *Asian J. Soc. Psychol.* 2, 1–19. doi: 10.1111/1467-839X.00023

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., et al. (2020). Crowdsourcing hypothesis tests: making transparent how design choices shape research results. *Psychol. Bull.* 146, 451–479. doi: 10.1037/bul0000220

Letiche, H., De Loo, I., Lowe, A., and Yates, D. (2023). Meeting the research(er) and the researched halfway. *Crit. Persp. Account.* 94:102452. doi: 10.1016/j.cpa.2022.102452

Mauthner, N. S. (2024). "Un/re-making method: knowing/enacting posthumanist performative social research methods through 'diffractive genealogies' and 'metaphysical practices" in Methods and genealogies of new materialisms, ed. F. Colman and TuinI. van der (Edinburgh: Edinburgh University Press), 186–211.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* 73, 719–748. doi: 10.1146/annurev-psych-020821-114157

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science* 349, aac4716. doi: 10.1126/science.aac4716

Pettigrew, T. F. (2018). The emergence of contextual social psychology. *Personal. Soc. Psychol. Bull.* 44, 963–971. doi: 10.1177/0146167218756033

Popper, K. (2002). Realismus und das Ziel der Wissenschaft. Tübingen: Mohr Siebeck.

Pothos, E. M., and Busemeyer, J. R. (2022). Quantum cognition. *Ann. Rev. Psychol.* 73, 749–778. doi: 10.1146/annurev-psych-033020-123501

Salter, P. S., Adams, G., and Perez, M. J. (2018). Racism in the structure of everyday worlds: a cultural-psychological perspective. *Curr. Dir. Psychol. Sci.* 27, 150–155. doi: 10.1177/0963721417724239

Scholz, J. (2013). "The possibility of a quantitative queer psychology" in Queering paradigms III - queer impact and practices. eds. K. O'Mara and L. Morrish (Oxford: Peter Lang), 239–258.

Scholz, J. (2018). Agential Realism als Basis queer(end)er Experimentalpsychologie. Eine wissenschaftstheoretische Auseinandersetzung. Wiesbaden: Springer.

Schrader, A. (2012). Haunted measurements. *Differences* 23, 119–160. doi: 10.1215/10407391-1892916

Shotter, J. (2014a). Agential realism, social constructionism, and our living relations to our surroundings: sensing similarities rather than seeing patterns. *Theory Psychol.* 24, 305–325. doi: 10.1177/0959354313514144

Shotter, J. (2014b). From within the thick of it: human beings doing being human in languaged worlds. *Theory Psychol.* 24, 592–605. doi: 10.1177/0959354314541984

Skinner-Dorkenoo, A. L., George, M., Wages, J. E., Sánchez, S., and Perry, S. P. (2023). A systemic approach to the psychology of racial bias within individuals and society. *Nat. Rev. Psychol.* 2, 392–406. doi: 10.1038/s44159-023-00190-z

Spivak, G. C. (1988). "Can the subaltern speak?" in Marxism and the interpretation of culture. eds. C. Nelson and L. Grossberg (Urbana, IL: University of Illinois Press), 271–313.

Teo, T. (2008). From speculation to epistemological violence in psychology: a critical-hermeneutic reconstruction. *Theory Psychol.* 18, 47–67. doi: 10.1177/0959354307086922

Teo, T. (2010). What is epistemological violence in the empirical social sciences. *Soc. Pers. Psychol. Comp.* 4, 295–303. doi: 10.1111/j.1751-9004.2010.00265.x

Tobias-Renstrøm, S., and Køppe, S. (2020). Karen Barad, psychology, and subject models: why we need to take experience seriously. *Theory Psychol.* 30, 638–656. doi: 10.1177/0959354320903089

Trout, J. D. (2001). "Measurement" in A companion to the philosophy of science. ed. W. H. Newton-Smith (Malden: Blackwell), 265–276.

Uher, J. (2021). Psychometrics is not measurement: unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theoret. Philos. Psychol.* 41, 58–84. doi: 10.1037/teo0000176

Uher, J. (2022). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: a rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., and Reinero, D. A. (2016a). Contextual sensitivity in scientific reproducibility. *PNAS* 113, 6454–6459. doi: 10.1073/pnas.1521897113

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., and Reinero, D. A. (2016b). Reply to Inbar: contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *PNAS* 113, E4935–E4936. doi: 10.1073/pnas.1609700113

Weber, E. U., Constantino, S. M., and Schlüter, M. (2023). Embedding cognition: judgment and choice in an interdependent and dynamic world. *Curr. Dir. Psychol. Sci.* 32, 328–336. doi: 10.1177/09637214231159282

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 11 April 2025
DOI 10.3389/fpsyg.2024.1397553



OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
Lucas B. Mazur,
Jagiellonian University, Poland
Jörg-Henrik Heine,
Ludwig Maximilian University of Munich, Germany

*CORRESPONDENCE
Christof Kuhbandner
✉ christof.kuhbandner@ur.de

RECEIVED 08 March 2024
ACCEPTED 04 November 2024
PUBLISHED 11 April 2025

CITATION
Kuhbandner C and Mayrhofer R (2025) The hidden complexity of the simple world of basic experimental psychology: the principal and practical limits of gaining psychological knowledge using the experimental method. *Front. Psychol.* 15:1397553. doi: 10.3389/fpsyg.2024.1397553



# The hidden complexity of the simple world of basic experimental psychology: the principal and practical limits of gaining psychological knowledge using the experimental method

Christof Kuhbandner* and Roland Mayrhofer

Department of Psychology, University of Regensburg, Regensburg, Germany

Basic experimental research in psychology is based on the assumption that law-like behavior can be observed if the complexity of the human psyche is reduced by the creation of experimental settings in which simple psychical phenomena occur which reflect the effect of an isolated psychological mechanism. However, we show that this assumption does not hold for many phenomena studied in basic experimental psychology because even phenomena that are regarded as simple and fully controllable often fluctuate unpredictably as a function of unintentionally chosen details of the experimental setting. The reason is that in a complex system like the human psyche, even minimal, and from the perspective of the investigated research question irrelevant, differences in the experimental setting can build up to large unsystematic effects. Law-like behavior in experiments could only occur if truly low-level mechanisms were studied in a truly isolated way. However, this is often not the case in current experimental research. One problem is that often fuzzy theoretical terms are used which only give the impression that low-level mechanisms are being investigated, although in reality the complexity of the human psyche is unintentionally brought on board. Another problem is that, unlike in the natural sciences, the mechanisms of the human psyche can only be isolated from each other to a limited extent because the human psyche always reacts as a whole system. If such problems could be overcome, meaningful knowledge could be gained through experimental psychological research. However, the knowledge gained is very limited in terms of its explanatory power for human behavior, as it is only helpful for understanding a very specific aspect of behavior, namely the mechanistic functioning of isolated low-level mechanisms. When it comes to understanding motivated behavior in real life, knowledge about the non-mechanistic functioning of the higher levels of the human psyche is necessary, but this knowledge cannot be gained through the experimental method.

KEYWORDS

**experimental psychology, experimental method, methodology, replication crisis, epistemology, complexity**

## 1 Introduction

One of the defining elements of any science is the method used to attempt to gain knowledge about the subject of research. A currently widespread methodological approach to gaining knowledge in the field of psychological science is the experiment, a method that has proven to be very fruitful in the field of natural sciences.[1] Particularly in the field of basic experimental psychology, there is a prevailing conviction that general laws of the human psyche can be established by means of the experimental method, similar to the natural sciences. The aim of this article is to critically examine this conviction.

## 2 The goal of science

It is a commonly shared view that the goal of science is to develop knowledge that allows us to predict which phenomena will occur if certain conditions are present. The most fundamental prerequisite for the development of such knowledge is that regularities are observed when a phenomenon is explored. In the most basic sense, "regularity" means that a certain observation that is made when a certain condition is present is observed again when the same condition is present again. Only if this is the case, knowledge about the phenomenon can be gained in the sense that theories can be developed which allow to predict what will happen if certain conditions are present.

However, to understand the great success of science, it is important to realize that science strives to establish theories about the existence of a certain form of regularity. For example, the Encyclopedia Britannica defines "science" basically as follows:[2]

> "In general, a science involves a pursuit of knowledge covering general truths or the operations of fundamental laws."

According to such definitions, the ultimate goal of science is not to establish theories that predict the occurrence of phenomena at the level of a singular object, but to establish *general* theories that predict the occurrence of phenomena for many different objects. For instance, the goal of physics as a science is not to establish a theory that predicts what movement is observed when a specific apple falls from a specific tree, but to establish a *general* theory that describes the falling movement of any object anywhere on Earth. This goal is achieved by postulating a certain cause-and-effect mechanism at the level of a property that is shared by many different objects. The objects to which the theory applies can nevertheless all be unique because they can differ in other object properties about which the theory makes no statements.[3]

## 3 The experimental method as the basis for the successful establishment of general theories in the natural sciences

With regard to the goal of establishing general theories, impressive successes have been achieved in the field of natural sciences. For instance, in physics, several universal laws were established that appear to exactly predict for any object anywhere in the universe what will be observed if a certain condition is present, such as the law of thermodynamics or the four laws of force (Ulanowicz, 2018).

This success was by and large made possible by using a very specific method to empirically test the validity of a proposed general theory: the experimental method. The use of this method was necessitated by an epistemic problem that arises when attempting to empirically test a general theory. To examine whether the predictions of a general theory correctly describe the occurrence of phenomena, it is necessary to explore whether all objects that have the property for which the theory formulates a cause-effect mechanism behave as predicted by the theory. However, objects not only have the specific property for which the theory under investigation makes a prediction, but also other properties on which other cause-effect mechanisms operate than that specified in the theory under investigation.

Accordingly, if one simply observed the behavior of objects in real life, one could not validate whether the predictions derived from a certain cause-effect mechanism correspond to the observations made, because the observed behavior is always determined by the interplay of all cause-effect mechanisms that simultaneously operate on the various object properties. Due to this fact, general theories that predict a certain cause-effect relationship cannot be empirically validated in real-life situations. An illustrative example is the law of gravitation. According to the law of gravitation, gravity accelerates every object at exactly the same rate so that heavy and light objects should fall at exactly the same speed. However, if one simply drops a feather and a steel ball in real life, one will observe that this is not the case, which seems to refute the law of gravitation. The reason why feathers and steel balls fall at different speeds in real life is that, in addition to gravity, there is a second influencing factor: air resistance.

This epistemic problem made it necessary to develop a method that allows the cause-effect mechanism specified by a specific general theory to be examined in isolation from all other simultaneously operating cause-effect mechanisms. And this is exactly what is achieved by the experimental method, which consists of deliberately manipulating the cause specified in the theory under investigation and measuring the resulting effect, while at the same time trying to eliminate the effects of all other additionally operating cause-effect mechanisms. An illustrative example is the way in which it could be empirically demonstrated that the law of gravitation makes correct predictions. This was made possible by the fact that an experimental setting was created in which objects were only influenced by gravity and no longer by air resistance, which was achieved by letting different objects fall in a vacuum. And indeed, under such conditions, feathers and steel balls fall at exactly the same speed, as predicted by the law of gravitation.

## 4 The adoption of this scientific logic in the field of psychological science

In view of the successes in establishing general theories by means of the experimental method in the field of natural sciences, the hope was

1 Here, the term "experimental method" always refers to the use of experiments as a scientific method for establishing general laws.

2 https://www.britannica.com/science/science

3 In the following, the term "general theory" always refers to theories that predict certain cause-effect relationships at the level of an object property which should apply to all objects that have this property.

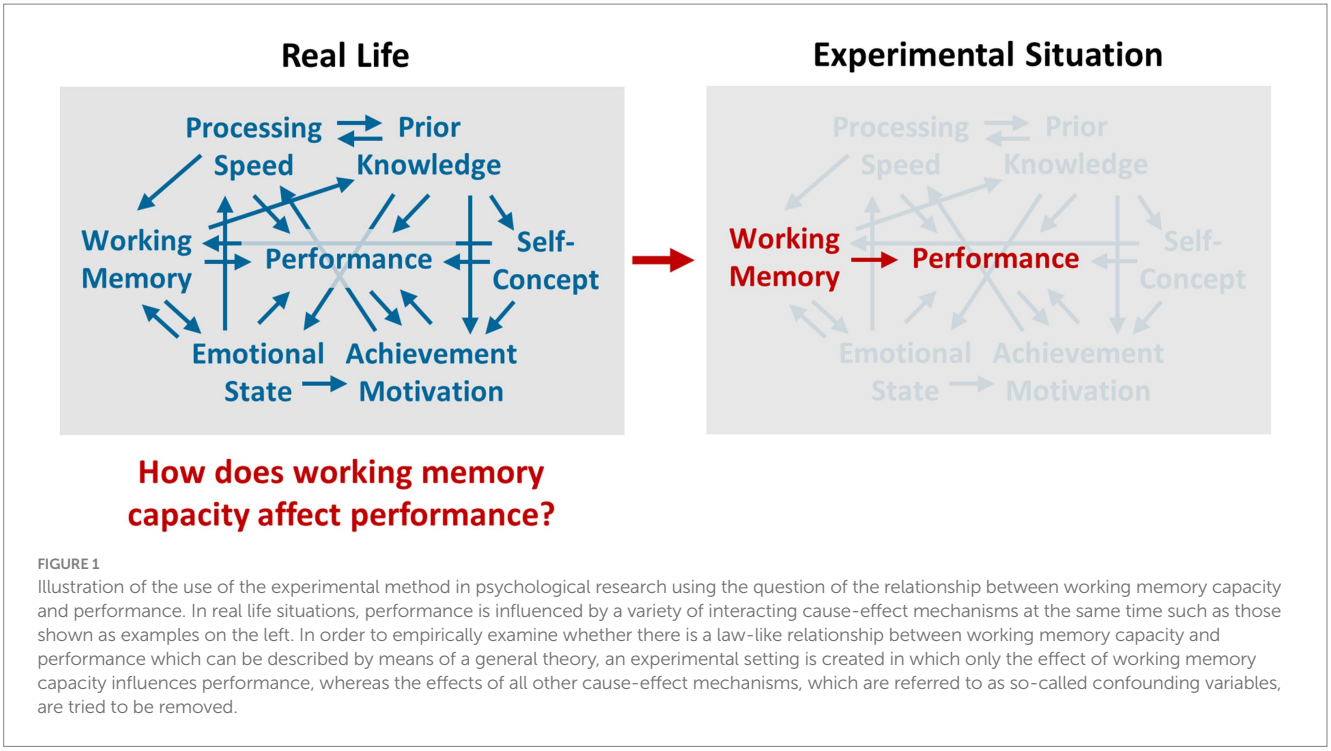



FIGURE 1
Illustration of the use of the experimental method in psychological research using the question of the relationship between working memory capacity and performance. In real life situations, performance is influenced by a variety of interacting cause-effect mechanisms at the same time such as those shown as examples on the left. In order to empirically examine whether there is a law-like relationship between working memory capacity and performance which can be described by means of a general theory, an experimental setting is created in which only the effect of working memory capacity influences performance, whereas the effects of all other cause-effect mechanisms, which are referred to as so-called confounding variables, are tried to be removed.

raised that the same scientific logic can be applied in the field of psychological research where the aim is to predict the occurrence of psychical phenomena,[4] and that comparable successes can be achieved (for an illustration of the use of the experimental method in psychological research, see Figure 1; for a description of the research history of the experimental method in the field of psychology, see Mandler, 2007).

Such a conviction is particularly common in the field of basic experimental psychological research, where attempts are made to gain knowledge about basic processes of the human psyche such as, for instance, perception or the storage of information. Characteristic of this field of research is the strong belief that law-like behavior is observed if the complexity of the human psyche is reduced by the creation of experimental settings in which simple psychical phenomena occur which reflect the effect of an isolated psychical mechanism. For instance, in an editorial of the journal *Experimental Psychology*, the editors describe the principles that characterize high-quality research as follows (Eder and Frings, 2018, p. 258):

> "First, it should be noted that experimentation is the 'golden standard' of scientific knowledge seeking. Experiments provide insight into cause and effect by systematic investigation of what outcome occurs when a particular factor or variable is manipulated. (…) A strong experiment gives great confidence in the inference of a causal relationship among variables."

And indeed, it is often claimed that it can be shown by means of basic experimental psychological research that certain psychical phenomena are governed by general laws. For instance, many articles and textbooks explain the course of forgetting of information stored in memory with recourse to a general law because there seems to be one retention function that describes the course of forgetting for many different types of memory as well as different memory contents (i.e., the power law of forgetting: the rate of decay slows with the passage of time; e.g., Rubin and Wenzel, 1996; Wixted and Ebbesen, 1991). Another example is the amount of information that can be held in working memory. It was proposed early on that there is a fixed number of items that humans can hold in working memory, with the suggestion that this number is 7 ± 2, which is frequently referred to as Miller's Law (Miller, 1956).

## 5 A first limit for the establishment of general theories: probabilistic instead of invariable cause-effect relationships

However, it became apparent that there are obviously limitations to describing psychical phenomena using general theories. In classical physics, it is the case that a cause always produces the same effect for all objects for which a theory is valid when all other cause-effect mechanisms operating on an object are excluded. That is, causes are invariably followed by their effects. However, as it turned out, such invariable patterns of causations are typically not observed in experimental psychological studies. There, the psychical phenomena that are expected to occur if a certain cause is present according to a postulated psychological theory do not occur invariably when the cause occurs, but only with a certain probability (e.g., Baumeister and Lau, 2024).

4 Following a suggestion by Uher (2021), this article uses the term "psychical" when referring to the phenomena that are explored, and the term "psychological" when referring to the means used for the exploration of a phenomenon.

A common explanation for the fact that only probabilistic rather than invariable cause-effect relationships are observed in experimental psychological studies is the high heterogeneity of psychical phenomena (Hitchcock, 2018). There are various psychological mechanisms in the human psyche, each of which trying to influence behavior in response to an event in its own way and according to its own standards. Since it is always the person's entire psyche with all its different mechanisms that reacts to an event, a specific psychological mechanism can be isolated from all other psychological mechanisms only in limited ways. To use an analogy: If a feather or a steal ball are placed in a physical vacuum, their usual way of reacting to physical forces will not change. But if one tried to put humans in a "psychological vacuum" in the sense that their inner psychological forces are eliminated (if that were even possible), then they would probably go mad.

For this reason, unlike in the natural sciences, the effects of cause-effect mechanisms that are not the focus of the theory under investigation cannot be completely excluded in experimental psychological studies. Indeed, this fact is also reflected in the quality standards that define the best possible way to conduct experiments in the field of psychology. For instance, in the already mentioned editorial of the journal *Experimental Psychology*, the editors describe the best possible way to conduct experiments as follows (Eder and Frings, 2018, p. 258):

> "The design of experimental research should be guided by the *max-con-min* principle: *maximize* the systematic variance of the experimental variables under scrutiny; *control* systematic error variance (or "bias") induced by confounding variables; and *minimize* random error variance induced by random variables."

Interestingly, a third category of effects is introduced in addition to the effect of the investigated cause-effect mechanism that is deliberately manipulated and the effects of the cause-effect mechanisms that are tried to be eliminated: there are obviously further effects (i.e., "random variables") whose causes are unknown, and which thus unpredictably bias the observed effects of the investigated cause-effect mechanism.

From a methodological perspective, this is often not seen as a major problem. The argument is that as long as the unknown cause-effect mechanisms are independent of each other and vary randomly and unsystematically across situations and persons, a specific mechanism can nevertheless be isolated by collecting many observations and averaging across the observations. By doing so, only the mechanism one is interested in causes systematic effects at the level of the averaged observations while the unknown mechanisms cause unsystematic random effects which level each other out. In fact, this is the research logic that almost all experimental studies follow today: a theoretically postulated cause-and-effect mechanism is examined at the level of averaged observations in a sample of individual persons that is supposed to be representative of the population about which the theory makes statements.

However, such a research logic has an often-overlooked consequence regarding the type of phenomena about which knowledge is generated. One often encounters the belief that this type of research would provide knowledge about the occurrence of psychical phenomena at the level of individual persons. However, this belief is actually misleading because the level of observation is not individual persons but averaged observations across individuals. Drawing conclusions from cause-effect relationships observed at the level of averaged observations across individual persons about the existence of cause-effect relationships at the level of individual persons would only make sense if a premise were fulfilled: the individual persons must be homogeneous in terms of the psychological structures and processes producing the observed phenomenon. In this case, how people react on average when they encounter an event would be informative for how an individual person reacts to the event, because the same pattern as observed on average at the group level would show up when an individual person repeatedly encounters the event.

However, numerous research findings call this premise into question, suggesting that heterogeneity instead of homogeneity is a defining characteristic of the functioning of the human psyche (e.g., Richters, 2021). Indeed, what distinguishes the human psyche is precisely that genetically underdetermined psychical structures and processes exist whose functioning parameters are determined by the experiences made in the idiosyncratic physical, social, and cultural environment. Such biographically determined individual adaptation processes can be found right down to the neuronal level. For instance, the experience-dependent elimination of neurons and synapses ("pruning") is regarded as one of the most important developmental mechanisms that enables the brain to adapt to the demands of the individual environment (e.g., Sakai, 2020).

As can be mathematically shown (i.e., the ergodic theorems), strict conditions would actually have to be met in order to transfer cause-effect relationships observed at the level of averaged observations across persons to the level of an individual person. However, these conditions are almost never checked and actually rarely met in psychological research (Molenaar and Campbell, 2009). This fact is particularly evident in experimental studies in which the behavior of people in real life is studied. For example, it is a common method in the field of educational science to investigate the effect of a learning method in an experiment in which the average performance in a group of people using the learning method is compared with the average performance in another group of people not using the learning method. However, individual performance varies around the averaged performance of the group, which means that the learning method gives some people a stronger advantage, while others have no advantage or possibly even a disadvantage. And since the average performance of the group does not provide any information about whether an individual person's performance is above or below the average performance, the observation that persons on average benefit from a certain learning method does not allow conclusions to be drawn as to whether the learning method is also effective for a particular individual person.

Given this fact, it is worth pointing out that classic definitions of the field of psychological research actually contain a misleading inaccuracy. For instance, according to the definition of the American Psychological Association, "psychology is the study of the mind and behavior".[5] Such definitions give the impression that psychological science studies mind and behavior at the level of individuals. However, since most experimental psychological studies actually

5 https://www.apa.org/support/about-apa

explore psychical phenomena at the level of averaged observations across individuals, a more adequate definition would actually have to include the addition "psychological science is the study of the mind and behavior *at the level of averaged observations across individuals*."

## 6 A possible second limit for the establishment of general theories: the occurrence of an irresolvable uncertainty in psychological research findings

The previous explanations show that there is a fundamental limit to the attempt to establish general theories of the functioning of the human psyche, namely that only probabilistic cause-effect relationships at the level of averaged observations across individuals can be empirically demonstrated. However, several recent studies suggest that the limitations are even more fundamental. The probabilistic cause-effect relationships observed in a specific study should be replicated when the same study is carried out again. However, as shown in several recent studies, this is not the case.

A first indication of a general replication problem emerged in a large-scale attempt to replicate 100 experimental and correlational studies published in high-ranking scientific psychological journals (Open Science Collaboration, 2015). While 97 % of the original studies had reported significant results, only 36 % of the replications had significant results. A similar picture emerged in a recent study where a text-based machine learning model was used to estimate the replication likelihood for more than 14,000 articles in six subfields of psychology published from 2000 to 2019 (Youyou et al., 2023a). The machine learning model was trained on the main texts of 388 manual replication studies in psychology that reported pass/fail replication outcomes to predict a paper's replicability based on the text in the manuscript. The results suggest that the mean likelihood of successful replication for a published psychological paper is only 0.42 (for criticisms, see Crockett et al., 2023; Mottelson and Kontogiorgos, 2023; for a reply to the criticisms, see Youyou et al., 2023b).

An initial reaction to the replication crisis from the psychological research community was the assumption that questionable research practices (e.g., Simmons et al., 2011), problematic incentive structures (e.g., Fanelli, 2010), and statistical misconceptions (e.g., Greenland et al., 2016) were responsible for the low replication rate, which led to various initiatives to improve the quality of research methods in psychology in order to increase the replication rate (e.g., Korbmacher et al., 2023). However, as shown in the above-mentioned study where the replication likelihood for psychological articles from 2000 to 2019 was estimated (Youyou et al., 2023a), the improvements in method-related and incentive-related problems had only a comparatively small impact. The average replication likelihood had decreased by approximately 10 % from 2000 and 2010 before the replication problem was brought to attention and before the various initiatives were launched. After that, the replication rate returned to the 2000 level and was still below 0.50 in 2019, suggesting that the reason for the problem of empirically demonstrating general regularities in psychology may be more fundamental than only the existence of questionable research practices and problematic incentive structures.

That this is indeed the case is shown by several recent studies which demonstrate that even exactly the same data set does not allow simple-sounding psychological questions to be empirically answered clearly and unambiguously. For instance, in a study by Silberzahn et al. (2018), 29 research teams were asked to empirically answer the question of whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players, based on exactly the same data set. The result pattern showed that the different analysis methods used by the different research teams did not converge. The estimated effect sizes ranged from 0.89 (less likely) to 2.93 (more likely) in odds-ratio units, and neither the teams' prior beliefs about the effect of interest nor their level of expertise nor the quality of the used methods readily explained the variation in the outcomes of the analyses. A similar finding was reported in a recent study by Breznau et al. (2022), where 73 independent research teams used exactly the same data set to empirically answer the question of whether more immigration will reduce public support for government provision of social policies. Instead of convergence, the results reported by the different research teams varied greatly, ranging from large negative to large positive effects of immigration on public support, and the variance in the obtained results was again not explained by the quality of research methods or the level of expertise.

These findings suggest that even when the problems of questionable research practices and biasing incentive structures are completely removed, and even when exactly the same data set is used when trying to answer a simple-sounding psychological question, it is impossible to establish reliable general theories. Instead, it seems, that there is an uncertainty in psychical phenomena that hampers attempts to establish general theories about the functioning of the human psyche.

## 7 Is basic experimental psychological research also affected by the occurrence of an irresolvable uncertainty?

In the two mentioned studies on the occurrence of an irresolvable uncertainty, psychical phenomena occurring in real-life were examined, which means that numerous mechanisms of the human psyche interact in a variety of ways without any experimental control. It could therefore be hoped that an irresolvable uncertainty will not occur in experiments in the field of basic experimental psychology, where simple psychical phenomena are investigated in artificial laboratory environments under carefully controlled conditions. And indeed, as already described above, this belief is very widespread in this field of research.

However, the results of the studies on the replicability of psychological studies suggest that the problem of uncertainty does also affect experimental studies, thus casting initial doubt on the assumption that the uncertainty observed in psychological studies may disappear in basic experimental psychology. If the use of the experimental method is associated with a lower uncertainty in the observed findings, the replication likelihood should be higher for experimental compared to non-experimental psychological studies. However, in the above-mentioned study (Youyou et al., 2023a) where the replication likelihood for more than 14,000 published psychological studies was estimated, the opposite was observed: the replication likelihood was lower for experimental studies than for non-experimental studies, a finding that was observed for all six subfields of psychology.

This finding suggests that there is a peculiarity in the functioning of the human psyche which entails that even when apparently simple psychical phenomena are explored in artificial laboratory environments under carefully controlled conditions, no law-like behavior can be observed. As already briefly mentioned, a characteristic of the human psyche is that it is a system which consists of numerous components that mutually influence each other and collectively shape the observed psychical phenomena. This type of organization has fundamental consequences for the occurrence of regularity in behavior.

As shown in the domain of chaos research, even if all components of such a system function in a strictly deterministic manner, it is impossible to predict what behavior the system will exhibit when it encounters certain conditions. The reason is that the smallest differences in the initial conditions can build up and alter the behavior of the system, which makes the behavior of the system unpredictable, a phenomenon called deterministic chaotic behavior (for a comprehensive description, see Prigogine and Stengers, 1997). An illustrative example is a pendulum that swings back and forth over two magnets. Unless the pendulum is not released directly over one of the two magnets, it is impossible to predict over which magnet the pendulum will come to rest, because minimal and no longer measurable shifts in the starting position of the pendulum can lead to different end positions. The phenomenon of chaotic behavior has entered everyday language in figurative form of the so-called "butterfly effect," which refers to the hypothetical assumption that large-scale phenomena such as tornados can be influenced by such small differences in the initial conditions as the flapping of a butterfly's wings (for a discussion, see Pielke et al., 2024).

The occurrence of chaotic behavior in systems consisting of mutually influencing components suggests that the assumption that law-like behavior can be observed when simple psychical phenomena are explored in highly controlled experimental settings may not be true. Given that in such systems as the human psyche, minimal differences in the initial conditions can lead to large and unpredictable differences in the observed behavior, it could be that even when exploring apparent simple psychical phenomena in an experimental setting with careful control of unwanted cause-effect mechanisms, still an irresolvable uncertainty occurs because the observed phenomena unpredictably vary as a function of minimal, and from the perspective of the investigated research question irrelevant, details of the experimental setting.

A closer look at the inner organization of the human psyche reveals another possible reason why even the apparently simple psychical phenomena that are explored in the field of basic experimental psychology may not show regularities that can be described by general laws. What distinguishes the human psyche from mechanistic systems such as a pendulum swinging over two magnets is that the inner components not only mutually influence each other. In addition, the inner components are additionally organized within a multi-layered structure of ascending levels of increasing organizational complexity. The special characteristic of such complex systems[6] is that at the higher levels of organization, novel phenomena with novel properties emerge that do not exist at the lower levels. The emergent phenomena on the higher levels in turn influence the functioning of the mechanisms on the lower levels in order to make them subserve the objectives pursued at the higher levels (e.g., Feinberg and Mallatt, 2020).

An illustrative example is the phenomenon of the experience of emotions. One of the most common definitions defines emotions as episodes of interrelated, synchronized changes in the states of all five organismic subsystems (cognitive, neurophysiological, motivational, motor expression, subjective feeling) in response to the evaluation of an external or internal stimulus event as relevant to major objectives of the organism (Scherer, 2005). Emotions therefore emerge on a higher organizational level in the sense of an organizational structure which provides various cross-system reaction patterns, and the mechanisms on the lower levels change depending on which emotion is currently experienced on the higher level.

The example of the higher-level mechanism of experiencing emotions illustrates why it makes no sense to postulate that the functioning of a low-level mechanism can be described by a general law if the mechanism is an integrative part of a complexly organized system. Since in such systems the concrete operating principles of the lower-level mechanisms are determined by the phenomena occurring on the higher levels, there is simply no general operating principle that could be described by a general law. For instance, it makes no sense to claim that the functioning of iconic memory, which is considered one of the basic cognitive processes of the human psyche, can be described by a general law because studies show that the properties of iconic memory vary as a function of the emotions currently experienced at the higher level of organization (e.g., Kuhbandner et al., 2011a,b). And since people are in a certain emotional state at every time point in their lives, it makes no sense to claim that the properties of iconic memory can be explained by a general law.

One could still hope that it may at least be possible to observe general regularities for certain interactions between low-level and high-level mechanisms. For instance, it could be that although the properties of iconic memory cannot be described in the form of a general law, the respective functioning in a certain emotional state can be described in the form of a general law. However, this hope is dashed by another peculiarity of the human psyche. What characterizes the human psyche is not only that its components are organized within a multi-layered structure of ascending levels of increasing organizational complexity. As already briefly mentioned above, what makes the human psyche special is that the psychical structures and processes at the higher levels are not genetically fixed but idiosyncratically developed based on the physical, social, and cultural environment of a particular individual. The consequence is that there is no general regularity that can be described by means of general theories because the functioning of lower-level mechanisms changes as a function of higher-level mechanisms which do not function in mechanistic ways but in idiosyncratic ways.

An illustrative example is the attempt to establish the functioning of emotions empirically. Initially, emotion research was guided by the hypothesis that each emotion has its own essence, that is that each emotion can be described by a separate mechanism that is specific to that emotion. If this were the case, each emotion would follow a certain general law, which could then be empirically proven. However, after hundreds of studies, the picture that has emerged is completely different. Both at the level of the facial, the cognitive, the motivational,

6 The term "complex system" is not used uniformly in the literature (for an overview, see Ladyman et al., 2013). In the present article, we use this term as an umbrella term to describe systems in which emergent phenomena occur from a collection of interacting parts.

the physiological, and the neuronal level, tremendous variation both within and across emotional categories is observed across studies, even when the same methods, stimuli, and sampling from the same population of participants were used, a pattern of finding that has been summarized in an overview in the statement "variation is the norm is a fair summary of the experimental literature on emotion" (Barrett and Westlin, 2021). In view of this variability, a new theoretical framework has been established that assumes that emotions have no essence but are categories of variable instances that vary from context to context depending on what has been functional according to the past experiences of a person (Barrett, 2013).

## 8 Is there a "butterfly effect" in basic experimental psychological experiments?

The described characteristics of the functioning of the human psyche suggest that a previously overlooked problem could exist in the field of basic experimental psychology. It could be that even when examining apparently very simple psychical phenomena under apparently highly controlled laboratory conditions, no regularity in behavior can be observed. Although attempts are made to tailor the experimental situation as closely as possible to the cause-effect mechanism under investigation, there are numerous details of the experimental situation which are often unintentionally chosen because they appear to be irrelevant for the mechanism under investigation (e.g., the concrete color of stimuli, the concrete spatial and sequential arrangement of stimuli, the current affective state of a specific subject). However, because in complex systems such as the human psyche, even minimal, and from the perspective of the investigated research question irrelevant, details of the experimental situation or the internal state of the participants can have large and unpredictable effects, the effect observed in a particular experiment may actually not reflect a generalizable effect of the cause-effect mechanism that is purportedly investigated, but actually the effects of minor details that unintentionally occurred in this specific experiment. In particular, even if one tries to explicitly control the effect of such minor details, this may not necessarily solve this problem if the mechanism of interest actually systematically varies as a function of these details.

A look at various research paradigms in the field of basic experimental psychology shows that it indeed often turns out that initially obtained findings actually depend on minor details that were unintentionally chosen in the initial experiments. For instance, in research on visual memory, the colors of visual objects are typically unintentionally chosen by experimenters. However, as shown in a series of studies, basic processes such as color-form binding are not uniform processes that work the same for all types of colors. Instead, red colors are particularly strongly bound whereas green colors are particularly weakly bound (Kuhbandner et al., 2015a).

Such effects of the occurrence of uncertainty due to the use of different types of stimuli have led to some of the initially postulated laws of the human psyche proving to be untenable. For example, Miller's law on the capacity of working memory, mentioned at the beginning, was in view of numerous contradictory findings described as "the legend of the magical number seven" (Cowan et al., 2007), and replaced by the "magical number four," which seemed to better describe the regularities observed across experiments (Cowan, 2010). However, meanwhile the variation in the observed findings is so huge that neither the magic number seven nor the magic number four can satisfactorily describe the psychical phenomena occurring in studies on working memory. Instead, it was suggested to abandon the theory of a fixed capacity in favor of theories that postulate that the quantity of items that can be held in working memory depends on the precision of the stored representations, with humans being able to flexibly trade between quantity and precision depending on context such as the currently experienced emotions (e.g., Spachtholz et al., 2014).

The case that further research reveals that initially obtained effects turn out to be effects that are actually tied to minor details of the experimental situation is found not only at the level of the stimuli chosen in an initial study, but also at the level of the response format chosen. For instance, in a highly cited study on the capacity of visual long-term memory, a remarkably high ability was observed to discriminate previously seen objects from highly similar new objects, which led the authors to conclude that visual long-term memory has a massive storage capacity for object details (Brady et al., 2008). However, in subsequent research, it turned out that this ability strongly varies as a function of subtle details of the test used. Performance is remarkably high when a test is used where the object previously seen and the new object are presented simultaneously on the screen (two-alternative forced-choice recognition test), but not when a test is used where the two objects are shown individually on separate screens (old-new recognition test; Cunningham et al., 2015).

Complicating matters even further, it turned out that even when consistently using two-alternative forced-choice recognition tests, a convergent result pattern is not necessarily observed. An example is the research on the phenomenon of recognition without awareness. An initial study showed that when testing recognition for highly complex visual stimuli with a two-alternative forced choice recognition test, recognition performance was highly accurate although the subjects reported that they had the feeling of being unable to remember the stimuli (Voss et al., 2008). However, another research team was not able to replicate this finding although the original study was reproduced as closely as possible (i.e., the same stimulus set, the same stimulus presentation times, etc.), concluding that recognition without awareness is an elusive phenomenon (Jeneson et al., 2010). As it turned out, the reason of this inconsistency across experiments was a slight variation in the way the subjects were instructed, encouraging subjects to guess in one case and to respond more confident in the other case (Voss and Paller, 2010).

In addition to effects of minor details of the experimental setting used in a particular study, further effects arise from minor details of the environment in which a particular experiment is carried out. For instance, it has been shown that the results obtained with exactly the same experimental setting vary as a function of environmental factors such as the sex or the attire of the experimenter (e.g., Green et al., 2005) or the body posture of the subjects (e.g., Muehlhan et al., 2014), the latter being one of the main factors why findings obtained in non-imaging standard laboratory settings, where subjects typically perform experimental tasks sitting upright, are sometimes difficult to replicate in neuroimaging settings, where subjects typically perform experimental tasks lying in supine position.

However, even when exactly the same experimental task is performed by subjects in exactly the same laboratory setting, the obtained results can unsystematically vary across the participating

subjects. For instance, a prominent theory in the phenomenon area of attention is based on the assumption that attention can be allocated advantageously to specific objects in visual space, an ability called object-based attention (e.g., Watson and Kramer, 1999). However, it turned out that such effects were difficult to replicate. In a comprehensive attempt to resolve the confusion reported in previous studies (Pilz et al., 2012), it was on the one hand shown that the occurrence of such effects depends on minor details: object-based attention effects were only observed when the stimuli were arranged horizontally but not when they were arranged vertically. However, even worse, bootstrapping showed that object-based attention effects were not observed in all of the tested subjects but only in a small minority of the subjects. The authors conclude that computing averages across tested subjects in experiments may not be a suitable method to create theories of cognition and perception because the variation on the level of individual subjects has to be taken into account for a true understanding of how cognition and perception work.

Critically, the effects of minor details of the experimental situation that are initially erroneously viewed as irrelevant can be so subtle that a whole research community does not notice this, creating the wrong impression that there is a general theory although this is actually not the case. Such a case can occur when all of the studies conducted to test a general theory consistently use the same specific research method which actually represents a special case, without the research community noticing this fact. An example is the so-called motivational-compatibility effect, which assumes that for positive stimuli approach behavior is faster elicited than withdrawal behavior, whereas for negative stimuli withdrawal behavior is faster elicited than approach behavior. In countless studies in which subjects were presented random series of positive and negative stimuli and their response speed and frequency of errors for approach and avoidance behavior measured, such an effect seemed to occur consistently over and over again (for a meta-analysis, see Phaf et al., 2014).

However, it turned out that a hidden confounding variable at the level of a minor detail of the experimental situation was present in all of the studies of this type. As shown in a series of studies, in such experiments, strong valence-independent trial-by-trial effects are observed because switching from approach to withdrawal behavior is much easier than vice versa (Kuhbandner et al., 2015b). These asymmetrical switch costs strongly biased the observed effects on trials where the opposite behavior had to be shown in the previous trial, creating the illusion that there is a similar motivational-compatibility effect for both negative and positive stimuli. However, looking only at the trials that were not biased by these asymmetrical switch costs revealed that motivational-compatibility effects are actually largely absent for negative stimuli and much stronger for positive stimuli. It is also interesting that this study, despite being published in a topic-relevant journal (*Cognition and Emotion*), has not been cited once yet by the motivational-compatibility effect research community, and that, to our knowledge, no study has taken this fact into account to date, which demonstrates how immune research communities can be to methodological problems.

There is also the particularly problematic case where details of the experimental situation that later turn out to be relevant are initially considered so irrelevant that they are even not described in the methods section of studies. This case is particularly problematic because by reading just the methods section of a study one cannot conclude that these boundary conditions even may exist. A prominent example are the studies on the electrophysiological correlates of attention and memory by the famous EEG researcher Steven Luck (e.g., Luck, 2012). In his textbook on the event-related potential technique (Luck, 2014), there is a box at the end entitled "Keeping subjects happy," which describes how Luck treats his subjects in the laboratory: he tries to keep them happy by playing their favorite music throughout the whole experiment, noting that the music brought by his subjects included all kinds of genres from classical, pop, rock, metal, rap, country, electronic, ambient, and just about everything else imaginable. However, in his published scientific papers, this treatment of subjects is not mentioned. Obviously, he assumes that the affective state of a subject is irrelevant for the basic cognitive processes he is investigating.

However, as it turned out, basic cognitive processes and their electrophysiological correlates vary not only quantitatively but even qualitatively as a function of the affective state of a subject. For instance, when making participants happy by playing happy music and asking them to retrieve happy memories, visual objects are stored in the form of coherent object representations mediated by attention-related brain activities. By contrast, when making participants sad by playing sad music and asking them to retrieve sad memories, visual objects are stored in the form of independent feature representations mediated by preattentive brain activities (Spachtholz and Kuhbandner, 2017).

Finally, there is also the case where a theoretically postulated psychical mechanism is confirmed in numerous experiments, but it turns out that the regularity observed in the experiments has nothing to do with the postulated psychological mechanism itself but is actually the effect of a minor detail of the experimental situation, which was unintentionally kept the same in all experiments. An example is the research on the so-called anger-superiority hypothesis, according to which it is easier to detect angry faces than happy faces in a crowd of neutral ones. The possible existence of such an effect was initially suggested using pictures of real faces (Hansen and Hansen, 1988). In response to criticism that the observed effect might not be due to emotional causes but due to differences in low-level visual features, subsequent studies used line drawings of emotional faces that consisted of identical features that were just spatially aligned differently (e.g., using the same curved line for the mouth, only oriented upwards versus downwards; Oehman et al., 2001). However, there was still criticism that the presentation of upward or downward curved lines alone could be sufficient for the effect to occur, which was in fact shown in follow-up studies (Coelho et al., 2010).

This finding indicates that the postulated psychological mechanism of an alleged superiority of angry faces, which was initially viewed as empirically proven, was actually driven by an emotion-independent minimal detail of the experimental situation. More generally viewed, as shown in more recent meta-analyses, the research history of the anger-superiority hypothesis is another example where, as research into the phenomenon increases, it turns out that the initial hypothesis of a general pattern breaks down into many individual findings that can no longer be summarized in the form of a general theory. For example, the authors of a recently published meta-analysis on the electrophysiological correlates of the anger superiority effect conclude in the abstract (Liu et al., 2021, p. 1): "the mean effect size difference between angry and happy expressions was ns. N2pc effect sizes were moderated by sample age, number of trials, and nature of facial images used (schematic vs. real)

[…]. As such, possible adaptive advantages of biases in orienting toward both anger and happy expressions warrant consideration in revisions of related theory."

## 9 Possible solutions and resulting consequences

As shown in the previous section, the assumption prevalent in the field of basic experimental psychology that law-like behavior can be observed if the complexity of the human psyche is reduced by creating experimental settings in which apparently simple psychical phenomena occur under apparently highly controlled conditions is often not fulfilled. The reason is that a special property of complex systems such as the human psyche is ignored in current research practice, namely that minor and, from the perspective of the investigated research question, irrelevant details of the experimental situation or the internal state of the participants can produce large effects. This leads to the accumulation of many individual experimental findings which, however, do not contribute to a cumulative acquisition of knowledge due to the occurrence of an unsystematic variability across the individual findings.

The question of possible solutions to this problem seems to be at first glance easy to answer: law-like behavior in experiments can only occur if a postulated cause-effect mechanism is studied in a truly isolated way. In this case, even the smallest differences that are irrelevant from the perspective of the investigated research question can no longer produce any effects. However, this necessary precondition for the possibility of the occurrence of law-like behavior is accompanied by fundamental consequences for the intention to explore the human psyche with the experimental method.

### 9.1 Consequences at the level of theory

The precondition that a cause-effect mechanism must be studied in a truly isolated way is accompanied by certain requirements at the level of the theoretical concepts based on which cause-effect relationships are formulated. As a starting point for working out these requirements, it is first necessary to clarify what exactly is meant by the term "concept". Building on this, it is then necessary to work out what special features theoretical concepts should have so that an experimentally isolatable cause-effect mechanism can be postulated based on them.

From a philosophy of science perspective, one fundamental assumption is that concepts are products of the human psyche, which allow humans to abstract from the abundance of internally representable entities. This abstraction is achieved by assigning entities that can actually be distinguished from each other to an overarching common concept, which defines a property that characterizes the set of entities assigned to the concept. An illustrative example is the concept "red," which is an overarching property that represents as a common concept all of the actually different hues that belong to this concept. Another example is the concept "intelligence," which is an overarching property that represents as a common concept the entirety of a person's problem-solving abilities. As the examples of the concepts "red" and "intelligence" illustrate, concepts can never be directly observed as such. Instead of seeing "red" or "intelligence," we can only ever see the individual referents (i.e., the currently perceived hue or the currently observed problem-solving ability) that we have agreed on belong to the concepts of "red" or "intelligence."

With regard to the question of what special features theoretical concepts should have so that an experimentally isolatable cause-effect mechanism can be formulated based on them, a straightforward requirement is that the referents of a concept must be precisely defined. If this is not the case, degrees of freedom arise with respect to the determination of the details of the experimental setting, which creates room for the occurrence of an irresolvable uncertainty. This requirement can be well illustrated by comparing the characteristics of everyday language terms and scientific terms, as done in the following quote from Bischof (2014, p. 37; translated by the authors):.

> When we talk about psychical matters in everyday life, we use everyday language. These terms are strange creatures: blurred fields of meaning, knotted associations of fragments of ideas that condense around a core and run out towards the edge without clear boundaries. It is easy to say what a 'mountain' is near the summit. But where does it end, where does the 'valley' begin? What is the minimum number of hairs a 'brush' must have? (...) Scientists sometimes make use of the words found in everyday language. They speak, for example, of 'power' or 'work' or 'performance'. But they subject the concepts that such words are supposed to denote to a rigorous definition. They nail down their exact referents and excludes everything else.

Problematically, the theoretical concepts used in psychological theories often do not do justice to the requirement that the referents of a concept must be precisely defined (for a recent discussion of this problem, see Hutmacher and Franz, 2024). Instead, to quote Norbert Bischof again,

> one often avoids clear definitions, relying on one's everyday feeling for language; the terms are left unpurified in their cloud of unclear connotations, and so that this is not noticed so quickly, at least the everyday expression is replaced by a technical term (Bischof, 2014, p. 38; translated by the authors).

By doing so, only the illusion is created that the concepts on which a psychological theory is based are precisely defined, although in reality a hidden universe of uncertainty is introduced.

However, the use of precisely defined theoretical concepts is not sufficient to enable a true isolation of cause-effect mechanisms in experiments. This can be illustrated using the example of the concept "intelligence." If one defines "intelligence" as the entirety of a person's problem-solving abilities, and if it were the case that all existing problem-solving abilities are known, then the concept would be absolutely precisely defined. However, if one were to formulate a cause-effect mechanism based on a concept such as "intelligence" and attempt to isolate this mechanism in an experiment, this would be an impossible undertaking.

The reason for this has to do with a special property of concepts. Concepts can abstract from the abundance of internally representable entities with a low or high degree of abstraction. At the lowest level of abstraction, the referents of a concept are entities that each are concretely perceivable at a given moment. An example is the concept "red" which refers to the group of perceivable colors with a specific hue. Such low-level concepts are characterized by an unidimensional

structure because each of the referents carries the property defined at the concept level completely within itself (e.g., unidimensional concepts).

At the higher levels of abstraction, the referents of a concept are not entities that each are concretely perceivable at a given moment, but other concepts that are located at a lower level of abstraction. This ability enables humans to flexibly represent the complexities of the world and the psyche at increasingly higher levels of abstraction with increasingly broader concepts. An example is the concept "intelligence." For instance, at a lower level of abstraction, verbal working memory abilities and visual working memory abilities can be distinguished because they are each based on independent mechanisms. These abilities can be represented at the next higher level as a joint entity by assigning them to the broader concept "working memory ability." At the next higher level, the referents of the concept "working memory ability" can be assigned to the broader concept of "fluid intelligence," which represents as a joint entity all abilities that share the common feature that they are independent of previously acquired knowledge. And finally, the referents of the concept "fluid intelligence" can be assigned to the broader concept of "intelligence," which represents as a joint entity the entirety of a person's abilities, including the abilities that depend on previously acquires knowledge. Such higher-level concepts are characterized by a multidimensional structure because each of the referents represents only a part of the property that is defined at the level of the higher-level concept.

With regard to the attempt to truly isolate cause-effects mechanisms in experiments, theoretical descriptions based on higher-level multidimensional concepts such as intelligence are problematic. Multidimensional concepts do not represent a concrete mechanism that may exist in reality. Instead, they are aggregates of different mechanisms that are actually each represented by their own concepts at a lower level of abstraction. For instance, the concept "working memory" does actually not represent a concrete mechanism. Instead, this concept summarizes the results of the separate systems of verbal working memory and visual working memory, which each function based on their own principles. Consequently, although multidimensional concepts such as "intelligence" can be precisely defined, they do not allow to exactly specify which mechanism should be isolated in a concrete experiment because different mechanisms are represented as a joint entity, which leads to the occurrence of an unresolvable uncertainty. Accordingly, a necessary precondition for the occurrence of law-like behavior in experiments is not only that the examined theoretical concepts are precisely defined but also that they are unidimensional low-level concepts.

Problematically, however, the use of broad multidimensional concepts is common in current basic experimental psychology. This creates the illusion that the same cause-effect mechanism is examined in different experiments, although actually different operationalizations of the same multidimensional concept were implemented. An illustrative example is the experimental research on "attention" and "working memory." There are hundreds of studies that are either framed under the theoretical term "attention" or the theoretical term "working memory," which gives the impression that there exist two independent low-level psychological mechanisms within the human psyche, each with its own independent mode of functioning. However, if one were to look at the definitions found in typical studies on "attention" and "working memory," one might come to the conclusion that these two terms have actually a strongly overlapping range of meaning. For instance, "working memory" is commonly defined as the mechanisms that hold the information currently most relevant for an ongoing behavior available for processing (e.g., Oberauer, 2019), and "attention" is commonly defined as the mechanisms that select, modulate, and sustain focus on information currently most relevant for an ongoing behavior (e.g., Chun et al., 2011). And if one were then to set out to explore the respective meanings more deeply, a whole universe of interconnected lower-level mechanisms would open up (for such an attempt, see, e.g., Oberauer, 2019), all of which would actually have to be described separately in a theoretically more fine-grained way if experimental psychological research is to be conducted in a meaningful way.

## 9.2 Principal limitations

As shown, it is a necessary precondition for the occurrence of law-like behavior in experiments that the explanatory concepts used in the examined theory are precisely defined unidimensional low-level concepts. This fact results in a fundamental limitation as to which types of psychical phenomena can be meaningfully investigated using the experimental method.

As already briefly mentioned, precisely the ability to build broad and abstract mental concepts that allow to represent the complexity of the world in a non-complex way is one of the central functional principles of the human psyche. In fact, it is exactly this ability that allows humans to show stable behavior in a situation where normally no stability occurs due to the occurrence of deterministic chaos. To establish order in this chaos, higher-level psychological mechanisms had to be established which operate on concepts that abstract from the vast number of details that are actually distinguishable on the lower levels of abstraction, but whose distinguishability is unimportant from the perspective of the acting person (for a detailed model, see, e.g., Tononi, 2012).

Accordingly, there is a first fundamental limitation: From the fact that law-like behavior in experiments can only occur if an investigated cause-effect relationship is based on precisely defined explanatory concepts with a low degree of abstraction, and from the fact that it is precisely the characteristic of higher-level mechanisms of the human psyche that they function based on fuzzily defined concepts with a high degree of abstraction, it follows that the higher-level mechanisms of the human psyche cannot be meaningfully investigated using the experimental method.

However, there is a second fundamental limitation preventing the occurrence of law-like behavior in experiments even when precisely defined low-level mechanisms are examined: the functioning of a low-level mechanism must not vary as a function of states at the higher level of the human psyche. As described above, if this is the case, it makes no sense to postulate that the functioning of a mechanism follows a general rule because there simply is no general rule. The ignoring of this fact often leads to the occurrence of unfruitful discussions in experimental psychology. An illustrative example is the history of research on the question of how the features of visual objects are stored in memory. In two simultaneously published papers, contrasting findings were observed. The findings of a study by Utochkin and Brady (2020) suggested that objects are stored as unbound feature representations. By

contrast, the findings of a study by Balaban et al. (2019) suggested that objects are stored as feature-bound object representations. A common reaction to such contradictory findings is to conclude that more research is needed to clarify which of the two possibilities is correct. However, a more fruitful research strategy that was not considered in either of the two studies is to investigate whether the way the features of visual objects are stored in memory depends on higher-level psychological mechanisms. And in fact, it was shown that the way the features of objects are stored in memory does not follow a general law but qualitatively varies as a function of the emotional state of observers (Spachtholz and Kuhbandner, 2017).

## 9.3 Practical limitations

In summary, it can therefore be said that only a very specific type of psychological mechanisms can be meaningfully investigated using the experimental method, namely low-level mechanisms that function independently of the higher-level mechanisms. It is disputed whether such mechanisms even exist in the human psyche. On the one hand, hundreds of studies claim to have shown that higher-level states such as beliefs, desires, emotions, motivations, intentions, and linguistic representations exert top-down influences on low-level perceptual mechanisms, suggesting that low-level mechanisms that function independently of the higher-level mechanisms do not exist. However, it has been argued that actually none of these studies provides compelling evidence for true top-down effects on perception (Firestone and Scholl, 2016), suggesting that such low-level mechanisms may exist.

However, even if low-level mechanisms exist in the human psyche that function independently of the higher-level mechanisms, there is an additional practical limitation: it is extremely difficult to create experimental situations in which psychical phenomena occur that exclusively reflect the effect of such a low-level mechanism. The reason is that the higher-level mechanisms of the psyche nevertheless influence behavior, even if the mechanism under investigation functions independently of these mechanisms. For example, subjects typically think about what is actually being investigated, how their performance compares to others, how they could improve their performance, or just what is for lunch, which brings additional effects into play that do not necessarily influence the functioning of the mechanism under investigation, but nevertheless influence the behavior observed in an experiment.

A recent study on the capacity of visual working memory shows that such effects even occur in very simple experimental settings (Laybourn et al., 2022). In that study, participants were asked to verbalize any feelings or thoughts they are experiencing while performing a standard visual working memory task where participants were asked to remember simple colored squares. The results showed that a variety of thoughts occurred that substantially varied across participants. For example, some participants perceived the task as meaningless, others perceived the task as a game, while still others perceived the task as an exam situation. Out of the 19 participants, six participants reported a change in motivation, stating for instance that the performance achieved became less and less important for them over time and that they just clicked somewhere on the screen, and three participants stated that they tried different strategies to improve performance. These findings show that even in very simple experimental situations, it cannot simply be assumed that exactly the same psychological mechanism is active in all participants. The authors themselves sum up this problem very well:

> "As researchers, we would like participants to be more like machines sometimes, so we can examine their "hardware" most accurately. However, it seems that human functioning is more complex" (p. 1602).

# 10 Consequences for the aim of gaining useful knowledge to explain human behavior by the experimental method

In summary, the present paper shows that there is a fundamental limit to understanding the functioning of the human psyche by means of the experimental method: law-like behavior can only occur in experiments when precisely defined low-level mechanisms are investigated that function completely independent of the higher-level mechanisms of the human psyche. This raises a fundamental question: to what extent can the experimental method be used to gain knowledge that is useful for explaining human behavior?

In order to answer this question, the term "behavior" needs to be broken down in more detail. A first necessary distinction concerns the distinction between the explanation of behavior shown *in laboratory settings* and behavior shown *in real life* (i.e., the so-called 'real-world or the lab'-dilemma, for a discussion, see Holleman et al., 2020). If human behavior in a laboratory setting is to be explained in which a psychical phenomenon occurs that reflects the effect of a truly low-level mechanism that is truly isolated from all other mechanisms of the human psyche, knowledge gained from experimental psychological research can be helpful. However, if the human behavior in real life is to be explained, knowledge gained from experimental psychology has no explanatory power because the behavior shown in real life is never solely determined by the isolated effect of a low-level mechanism. Instead, in real life the human psyche with all its mechanisms always reacts to situations as a whole, with situations being sometimes even actively created by the human psyche in the first place.

However, the psychological knowledge that can be gained by means of the experimental method is not completely irrelevant for the aim to explain the behavior of humans in real life as sometimes claimed (e.g., Debrouwere and Rosseel, 2022). In order to see this point, a further distinction is necessary with regard to the term "behavior": the distinction between the explanation of *mechanistic* behavior and *motivated* behavior. This distinction can be illustrated using the following instruction:

> "Dear reader, please raise your hand!"

Let us assume that you as a reader have actually raised your hand. If one wants to explain this behavior, one can first take a neuroscientific perspective. And from this perspective, one will come to the conclusion that the raising of the hand was caused by an activation of the area in the brain that controls the hand movement. And from this perspective, one might even find oneself thinking that this brain activation fully explains the behavior, because whenever this brain

activation is observed in a person, they always raise their hand. However, although this is a truly causal explanation, it has no explanatory power whatsoever with regard to the question of why someone raised their hand. The actual cause why you as a reader raised your hand was the instruction that we as authors gave, and which you understood and followed. And that we wrote this instruction was of course also caused by an activation of our brains. But again, this does not provide an explanation, because the idea to give such an instruction in our paper came at the end of a long chain of thoughts that have built up in us over many years. And whether you as a reader really raised your hand in response to this instruction depends on whether you were motivated to follow this instruction. And that, in turn, depends on the individual views, beliefs and values that have built up over the years on your higher levels of the human psyche.

Accordingly, when one aims to explain an observed behavior, such as raising a hand, there are two separate types of knowledge which are necessary to explain the behavior. On the one hand, knowledge is needed about the mechanisms which underly the *general* ability to mechanistically react to certain sensory experiences with certain motor responses, regardless of when and under what motivational circumstances the behavior is actually shown (i.e., mechanistic behavior). On the other hand, knowledge is needed about the mechanisms which motivate a particular person to actually exhibit in a particular situation the motor behavior of which they are potentially capable (i.e., motivated behavior). And with the experimental method, helpful knowledge can be gained for the explanation of mechanistic behavior, but not for the explanation of motivated behavior.

Accordingly, experimentally gained knowledge can be important to explain behavior in real life in the sense that someone must have the *general* ability to perform a certain behavior in order to be able to show this behavior as a response. However, if one wants to understand when and under what circumstances a person shows a behavior in real life, knowledge gained from experimental psychology is not helpful. In this case, the question is about why a person is motivated to show a certain behavior, a question that can only be answered based on knowledge about the non-mechanistic higher-level processes of the human psyche which give meaning and direction to a person's behavior in real life – knowledge that cannot be gained by means of the experimental method.

There is a final important point that needs to be made. Someone could come up with the idea that the occurrence of regular behavior in experiments can also be achieved by setting the states of the tested participants on all levels of the human psyche exactly the same, except for the specific mechanism being investigated. However, if this were at all possible (for a critical discussion, see, e.g., Smedslund, 2016), one would be introducing a hidden assumption about the functioning of the human psyche, namely that it is possible to generalize the functioning of higher-level mechanisms across different people.

However, it is exactly the opposite that constitutes the special characteristic of the higher-level mechanisms of the human psyche. There is no general rule as to how the complexity of the world should ideally be mapped into broad and fuzzy mental concepts on the higher levels. Instead, the optimal granularity with which the world is conceptualized varies idiosyncratically as a function of the current external and internal context and the historical, cultural, and biographical background of an individual observer. If one wants to understand the uniqueness of the human psyche, methods have to be used that take into account the idiosyncratic functioning of the human psyche (for an overview, see Salvatore and Valsiner, 2023). Otherwise, if one were to try to make all people the same in an experiment, one would actually take away exactly what makes humans different from inanimate objects: that humans can react to exactly the same physical situation in unique ways.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## Author contributions

CK: Conceptualization, Writing – original draft, Writing – review & editing. RM: Writing – review & editing.

## Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Balaban, H., Assaf, D., Arad Meir, M., and Luria, R. (2019). Different features of real-world objects are represented in a dependent manner in long-term memory. *J. Exp. Psychol. Gen.* 149, 1275–1293. doi: 10.1037/xge0000716

Barrett, L. F. (2013). Psychological construction: the darwinian approach to the science of emotion. *Emot. Rev.* 5, 379–389. doi: 10.1177/1754073913489753

Barrett, L. F., and Westlin, C. (2021). "Navigating the science of emotion" in Emotion measurement. ed. H. L. Meiselman. *2nd* ed (Duxford: Elsevier), 39–84.

Baumeister, R. F., and Lau, S. (2024). Why psychological scientists should disdain determinism. *Possibil. Stud. Soc.* 2, 282–302. doi: 10.1177/27538699241258002

Bischof, N. (2014). Psychologie: Ein Grundkurs für Anspruchsvolle. *2nd* Edn. Stuttgart: Kohlhammer.

Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceed. Natl. Acad. Sci. U. S. A.* 105, 14325–14329. doi: 10.1073/pnas.0803390105

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceed. Natl. Acad. Sci. U. S. A.* 119:e2203150119. doi: 10.1073/pnas.2203150119

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol.* 62, 73–101. doi: 10.1146/annurev.psych.093008.100427

Coelho, C. M., Cloete, S., and Wallis, G. (2010). The face-in-the-crowd effect: when angry faces are just cross(es). *J. Vis.* 10, 7.1–7.14. doi: 10.1167/10.1.7. PMID: 20143900

Cowan, N. (2010). The magical mystery four: how is working memory capacity limited, and why? *Curr. Dir. Psychol. Sci.* 19, 51–57. doi: 10.1177/0963721409359277

Cowan, N., Morey, C. C., and Chen, Z. (2007). "The legend of the magical number seven" in Tall tales about the mind & brain: Separating fact from fiction. ed. S. Della Sala (Oxford: Oxford University Press), 45–59.

Crockett, M. J., Bai, X., Kapoor, S., Messeri, L., and Narayanan, A. (2023). The limitations of machine learning models for predicting scientific replicability. *Proceed. Natl. Acad. Sci. U. S. A.* 120, 1–2. doi: 10.1073/pnas.2307596120

Cunningham, C. A., Yassa, M. A., and Egeth, H. E. (2015). Massive memory revisited: limitations on storage capacity for object details in visual long-term memory. *Learn. Mem.* 22, 563–566. doi: 10.1101/lm.039404.115

Debrouwere, S., and Rosseel, Y. (2022). The conceptual, cunning, and conclusive experiment in psychology. *Perspect. Psychol. Sci.* 17, 852–862. doi: 10.1177/17456916211026947

Eder, A. B., and Frings, C. (2018). What makes a quality journal? *Exp. Psychol.* 65, 257–262. doi: 10.1027/1618-3169/a000426

Elliot, A. J., and Niesta, D. (2008). Romantic red: Red enhances men's attraction to women. *J. Pers. Soc. Psychol.* 95, 1150–1164. doi: 10.1037/0022-3514.95.5.1150

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS One* 5:e10271. doi: 10.1371/journal.pone.0010271

Feinberg, T. E., and Mallatt, J. (2020). Phenomenal consciousness and emergence: eliminating the explanatory gap. *Front. Psychol.* 11:1041. doi: 10.3389/fpsyg.2020.01041

Firestone, C., and Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for top-down effects. *Behav. Brain Sci.* 39:e229. doi: 10.1017/S0140525X15000965

Green, R. J., Sandall, J. C., and Phelps, C. (2005). Effect of experimenter attire and sex on participant productivity. *Soc. Behav. Personal. Int. J.* 33, 125–132. doi: 10.2224/sbp.2005.33.2.125

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31, 337–350. doi: 10.1007/s10654-016-0149-3

Hansen, C. H., and Hansen, R. D. (1988). Finding the face in the crowd: an anger superiority effect. *J. Pers. Soc. Psychol.* 54, 917–924. doi: 10.1037/0022-3514.54.6.917

Hitchcock, C. (2018). "Probabilistic causation" in The Stanford encyclopedia of philosophy (Winter 2016 edition). eds. E. N. Zalta and U. Nodelman (Redwood City, CA: Stanford University Press).

Holleman, G. A., Hooge, I. T. C., Kemner, C., and Hessels, R. S. (2020). The 'real-world approach and its problems: a critique of the term ecological validity. *Front. Psychol.* 11:721. doi: 10.3389/fpsyg.2020.00721

Hutmacher, F., and Franz, D. J. (2024). Approaching psychology's current crises by exploring the vagueness of psychological concepts: recommendations for advancing the discipline. *Am. Psychol.* Advance online publication. doi: 10.1037/amp0001300

Jensen, A. R. (2002). Psychometric g: Definition and substantiation. In *The general factor of intelligence* eds. R. J. Sternberg and E. L. Grigorenko (Psychology Press: New York). 39–53.

Jeneson, A., Kirwan, C. B., and Squire, L. R. (2010). Recognition without awareness: an elusive phenomenon. *Learn. Mem.* 17, 454–459. doi: 10.1101/lm.1815010

Kanazawa, S. (2011). Intelligence and physical attractiveness. *Intelligence* 39, 7–14. doi: 10.1016/j.intell.2010.11.003

Kay, P., and Regier, T. (2003). Resolving the question of color naming universals. *PNAS* 100. 9085–9. doi: 10.1073/pnas.1532837100

Koenderink, J. J., van Doorn, A. J., and Braun, D. I. (2024). Warm, cool, and the colors. *J. Vis.* 24:5. doi: 10.1167/jov.24.7.5

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., et al. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Commun. Psychol.* 1:3. doi: 10.1038/s44271-023-00003-2

Kuhbandner, C. (2020). Real-world objects are represented in visual long-term memory both as unbound features and as bound objects. *Front. Psychol.* 11:580667. doi: 10.3389/fpsyg.2020.580667

Kuhbandner, C., Lichtenfeld, S., and Pekrun, R. (2011a). Always look on the broad side of life: happiness increases the breadth of sensory memory. *Emotion* 11, 958–964. doi: 10.1037/a0024075 PMID: 21859210

Kuhbandner, C., Spitzer, B., Lichtenfeld, S., and Pekrun, R. (2015a). Differential binding of colors to objects in memory: red and yellow stick better than blue and green. *Front. Psychol.* 6:231. doi: 10.3389/fpsyg.2015.00231

Kuhbandner, C., Spitzer, B., and Pekrun, R. (2011b). Read-out of emotional information from iconic memory: the longevity of threatening stimuli. *Psychol. Sci.* 22, 695–700. doi: 10.1177/0956797611406445

Kuhbandner, C., Vogel, C. M., and Lichtenfeld, S. (2015b). Switching from approach to withdrawal is easier than vice versa. *Cognit. Emot.* 29, 1168–1184. doi: 10.1080/02699931.2014.969197

Ladyman, J., Lambert, J., and Wiesner, K. (2013). What is a complex system? *Eur. J. Philos. Sci.* 3, 33–67. doi: 10.1007/s13194-012-0056-8

Laybourn, S., Frenzel, A. C., Constant, M., and Liesefeld, H. R. (2022). Unintended emotions in the laboratory: Emotions incidentally induced by a standard visual working memory task relate to task performance. *J. Exp. Psychol. Gen.* 151, 1591–1605. doi: 10.1037/xge0001147

Liu, Y., Wang, Y., Gozli, D. G., Xiang, Y. T., and Jackson, T. (2021). Current status of the anger superiority hypothesis: a meta-analytic review of N2pc studies. *Psychophysiology* 58:e13700. doi: 10.1111/psyp.13700

Luck, S. J. (2012). "Electrophysiological correlates of the focusing of attention within complex visual scenes: N2pc and related ERP components" in The Oxford handbook of event-related potential components. eds. E. S. Kappenman and S. J. Luck (Oxford: Oxford Academic).

Luck, S. J. (2014). An introduction to the event-related potential technique. *2nd* Edn. Cambridge, MA: MIT Press.

Mandler, G. (2007). A history of modern experimental psychology: From James and Wundt to cognitive science. Cambridge, MA: MIT Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Molenaar, P. C. M., and Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Curr. Dir. Psychol. Sci.* 18, 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Mottelson, A., and Kontogiorgos, D. (2023). Replicating replicability modeling of psychology papers. *PNAS Proceed. Natl. Acad. Sci. U. S. A.* 120, e2309496120–e2309496122. doi: 10.1073/pnas.2309496120

Muehlhan, M., Marxen, M., Landsiedel, J., Malberg, H., and Zaunseder, S. (2014). The effect of body posture on cognitive performance: a question of sleep quality. *Front. Hum. Neurosci.* 8:171. doi: 10.3389/fnhum.2014.00171

Oberauer, K. (2019). Working memory and attention - a conceptual analysis and review. *J. Cogn.* 2:36. doi: 10.5334/joc.58

Oehman, A., Lundqvist, D., and Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *J. Pers. Soc. Psychol.* 80, 381–396. doi: 10.1037/0022-3514.80.3.381

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, 1–8. doi: 10.1126/science.aac4716

Phaf, R. H., Mohr, S. E., Rotteveel, M., and Wicherts, J. M. (2014). Approach, avoidance, and affect: a meta-analysis of approach-avoidance tendencies in manual reaction time tasks. *Front. Psychol.* 5:378. doi: 10.3389/fpsyg.2014.00378

Pielke, R. A., Shen, B. W., and Zeng, X. (2024). The butterfly effect: can a butterfly in Brazil really cause a tornado in Texas? *Weatherwise* 77, 14–18. doi: 10.1080/00431672.2024.2329521

Pilz, K. S., Roggeveen, A. B., Creighton, S. E., Bennett, P. J., and Sekuler, A. B. (2012). How prevalent is object-based attention? *PLoS One* 7:e30693. doi: 10.1371/journal.pone.0030693

Prigogine, I., and Stengers, I. (1997). The end of certainty: Time, Chaos, and the new Laws of nature. New York, NY: Free Press.

Reeves, A., Fuller, H., and Fine, E. M. (2005). The role of attention in binding shape to color. *Vis. Res.* 45, 3343–3355. doi: 10.1016/j.visres.2005.07.041

Richters, J. E. (2021). Incredible utility: the lost causes and causal debris of psychological science. *Basic Appl. Soc. Psychol.* 43, 366–405. doi: 10.1080/01973533.2021.1979003

Rubin, D., and Wenzel, A. (1996). One hundred years of forgetting: a quantitative description of retention. *Psychol. Rev.* 103, 734–760. doi: 10.1037/0033-295X.103.4.734

Sakai, J. (2020). Core concept: how synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceed. Natl. Acad. Sci. U. S. A.* 117, 16096–16099. doi: 10.1073/pnas.2010281117

Salvatore, J., and Valsiner, (2023). Ten years of idiographic science. Greenwich, CT: Information Age Publishers.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Soc. Sci. Inf.* 44, 695–729. doi: 10.1177/0539018405058216

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., et al. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* 1, 337–356. doi: 10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Smedslund, J. (2016). Why psychology cannot be an empirical science. *Integr. Psychol. Behav. Sci.* 50, 185–195. doi: 10.1007/s12124-015-9339-x

Spachtholz, P., and Kuhbandner, C. (2017). Visual long-term memory is not unitary: flexible storage of visual information as features or objects as a function of affect. *Cognit. Affect. Behav. Neurosci.* 17, 1141–1150. doi: 10.3758/s13415-017-0538-4

Spachtholz, P., Kuhbandner, C., and Pekrun, R. (2014). Negative affect improves the quality of memories: trading capacity for precision in sensory and working memory. *J. Exp. Psychol. Gen.* 143, 1450–1456. doi: 10.1037/xge0000012

Tononi, G. (2012). The integrated information theory of consciousness: an updated account. *Arch. Ital. Biol.* 150, 56–90. doi: 10.4449/aib.v149i5.1388

Uher, J. (2021). Psychology's status as a science: peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integr. Psychol. Behav. Sci.* 55, 212–224. doi: 10.1007/s12124-020-09545-0

Uher, J. (2023). What Are Constructs? Ontological Nature, Epistemological Challenges, Theoretical Foundations and Key Sources of Misunderstandings and Confusions. *Psychol. Inq.* 34, 280–290. doi: 10.1080/1047840X.2023.2274384

Ulanowicz, R. E. (2018). "The universal laws of physics: inflated ontologies?" in God's Providence and randomness in nature. eds. R. J. Russell and J. M. Moritz (Conshohocken, PA: Templeton Press), 69–84.

Utochkin, I. S., and Brady, T. F. (2020). Independent storage of different features of real-world objects in long-term memory. *J. Exp. Psychol. Gen.* 149, 530–549. doi: 10.1037/xge0000664

Voss, J. L., Baym, C. L., and Paller, K. A. (2008). Accurate forced-choice recognition without awareness of memory retrieval. *Learn. Mem.* 15, 454–459. doi: 10.1101/lm.971208

Voss, J. L., and Paller, K. A. (2010). What makes recognition without awareness appear to be elusive? Strategic factors that influence the accuracy of guesses. *Learn. Mem.* 17, 460–468. doi: 10.1101/lm.1896010

Watson, S. E., and Kramer, A. F. (1999). Object-based visual selective attention and perceptual organization. *Percept. Psychophys.* 61, 31–49. doi: 10.3758/bf03211947

Witzel, C. (2019). Misconceptions about colour categories. *Rev. Philos. Psychol.* 10, 499–540. doi: 10.1007/s13164-018-0404-5

Wixted, J. T., and Ebbesen, E. B. (1991). On the form of forgetting. *Psychol. Sci.* 2, 409–415. doi: 10.1111/j.1467-9280.1991.tb00175.x

Youyou, W., Yang, Y., and Uzzi, B. (2023a). A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proceed. Natl. Acad. Sci. U. S. A.* 120:e2208863120. doi: 10.1073/pnas.2208863120

Youyou, W., Yang, Y., and Uzzi, B. (2023b). Reply to crocket et al. and Mottelson and Kontogiorgos: machine Learning's scientific significance and future impact on replicability research. *Proceed. Natl. Acad. Sci. U. S. A.* 120:e2308195120:e2308195120. doi: 10.1073/pnas.2308195120

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 26 June 2025
DOI 10.3389/fpsyg.2025.1534270



OPEN ACCESS

EDITED BY
Alessandro Giuliani,
National Institute of Health (ISS), Italy

REVIEWED BY
Moritz Dechamps,
Ludwig Maximilian University of Munich, Germany
Gianfranco Minati,
Italian Systems Society, Italy
Michele Luchetti,
Bielefeld University, Germany

*CORRESPONDENCE
Jana Uher
✉ mail@janauher.com

RECEIVED 25 November 2024
ACCEPTED 26 May 2025
PUBLISHED 26 June 2025

CITATION
Uher J (2025) Statistics is not measurement: The inbuilt semantics of psychometric scales and language-based models obscures crucial epistemic differences.
*Front. Psychol.* 16:1534270.
doi: 10.3389/fpsyg.2025.1534270



# Statistics is not measurement: The inbuilt semantics of psychometric scales and language-based models obscures crucial epistemic differences

Jana Uher*

School of Human Sciences, University of Greenwich, London, United Kingdom

This article provides a comprehensive critique of psychology's overreliance on statistical modelling at the expense of epistemologically grounded measurement processes. It highlights that statistics deals with structural relations in data regardless of what these data represent, whereas measurement establishes traceable empirical relations between the phenomena studied and the data representing information about them. These crucial epistemic differences are elaborated using Rosen's general model of measurement, involving the coherent modelling of the (1) objects of research, (2) data generation (encoding), (3) formal manipulation (e.g., statistical analysis) and (4) result interpretation regarding the objects studied (decoding). This system of interrelated modelling relations is shown to underlie metrologists' approaches for tackling the problem of epistemic circularity in physical measurement, illustrated in the special cases of measurement coordination and calibration. The article then explicates psychology's challenges for establishing genuine analogues of measurement, which arise from the peculiarities of its study phenomena (e.g., higher-order complexity, non-ergodicity) and language-based methods (e.g., inbuilt semantics). It demonstrates that psychometrics cannot establish coordinated and calibrated modelling relations, thus generating only pragmatic quantifications with predictive power but precluding epistemically justified inferences on the phenomena studied. This epistemic gap is often overlooked, however, because many psychologists mistake their methods' inbuilt semantics—thus, descriptions of their study phenomena (e.g., in rating scales, item variables, statistical models)—for the phenomena described. This blurs the epistemically necessary distinction between the phenomena studied and those used as means of investigation, thereby confusing ontological with epistemological concepts—psychologists' cardinal error. Therefore, many mistake judgements of verbal statements for measurements of the phenomena described and overlook that statistics can neither establish nor analyze a model's relations to the phenomena explored. The article elaborates epistemological and methodological fundamentals to establish coherent modelling relations between real and formal study system and to distinguish the epistemic components involved, considering psychology's peculiarities. It shows that epistemically justified inferences necessitate methods for analysing individuals' unrestricted verbal responses, now advanced through artificial intelligence

systems modelling natural language (e.g., NLP algorithms, LLMs). Their increasing use to generate standardised descriptions of study phenomena for rating scales and constructs, by contrast, will only perpetuate psychologists' cardinal error—and thus, psychology's crisis.

KEYWORDS

**measurement, psychometrics, large language models (LLMs), natural language processing (NLP), rating scales, modelling relation, semantics-syntax, metrology**

# 1 Statistics vs. measurement

Psychology cherishes its sophisticated 'measurement' and modelling techniques for enabling quantitative research—the hallmark of modern science. A closer look reveals, however, that only methods of statistical data analysis are well elaborated, which together with pertinent research designs (e.g., between-subjects) fill our books and journals on psychological research methods. This emphasis reflects the prevailing view that statistics constitutes psychology's approach for 'measuring' its non-observable study phenomena (e.g., in psychometrics). This assumption, however, is based on epistemic errors because statistics neither *is* measurement nor is statistics necessary for measurement.

## 1.1 Different scientific activities for different epistemic purposes

Measurement and measurement scales have been successfully developed in physics and metrology—the science of physical measurement and its application (JCGM100:2008, 2008, p. 2.2)—long before statistics was invented (Abran et al., 2012; Fisher, 2009; Uher, 2022b, 2023a). Measurement and statistics involve different scientific activities designed for different epistemic (knowledge-related) purposes.

*Measurement* requires traceable empirical interactions with the specific quantities to be measured in the phenomena and properties under study—the *measurands* (e.g., person A's body temperature but not A's body weight or volume; person B's duration of speaking in a specific situation). Epistemically justifiable inferences from observable indications of these empirical interactions back to the measurands require theoretical knowledge about both the object of research and the objects used as measuring instruments as well as their conceptualisation in a defined process structure within a realist framework (Mari et al., 2021; Schrödinger, 1964; von Neumann, 1955). Its empirical implementation necessitates unbroken documented connection chains that establish proportional (quantitative) relations of the results with both (1) the measurand's unknown quantity (e.g., A's body temperature; B's duration of speaking)—the principle of *data generation traceability*—and (2) a known reference quantity (e.g., international units). This reference is necessary to establish the results' quantitative meaning regarding the specific property studied (e.g., *how* warm or *how* long that is)—the principle of *numerical traceability* (Uher, 2018a, 2020b, 2021c,d, 2022a,b, 2023a). Process structures thus-established allow for deriving epistemically justified information about specific quantities that are assumed to exist in an object of research and for representing this information in sign systems that are unambiguously interpretable regarding those measurands (e.g., '$T_{Pers\_A}$ = 36.9°C'; '$d_{Pers\_B}$ = 16.2 mins/h').

*Statistics*, by contrast, enables probabilistic descriptions of what might happen as a consequence of complex, poorly understood and possibly random events and processes as well as of constraints that are set by stochastic boundaries (e.g., distribution curves). In data sets, statistical methods allow us to identify regularities beyond pure randomness, to group cases and compare groups by their parameters, to model and extrapolate patterns as well as to estimate error and uncertainty for justifying inferences from samples to distribution patterns in hypothetical populations (Romeijn, 2017). Statistics builds on theories that define the workings of the analytical operations performed (e.g., mathematical statistics, probability theory, item response theory). But it does not build on theories about the objects of research that scientists may aim to analyse for prevalences, differences and trends, and that may be as diverse as diseases, therapeutic treatments, behaviours, intellectual abilities, financial markets, policies and others. Statistics is mute about the specific phenomena and properties analysed (Strauch, 1976). That is, statistics concerns the analysis of data sets *regardless of what these data are meant to represent*. Therefore, it does not require a term denoting the specific quantity to be measured in the real study objects—the measurand. This may explain why most psychologists are unfamiliar with this basic term. Their focus on 'true scores' in statistical modelling obscures the epistemic distinction between the real quantity to be measured and the measurement results used to estimate it (Strom and Tabatadze, 2022).

Statistics, however, is fundamental to so-called psychological 'measurement'. Why?

## 1.2 Psychological 'measurement': Statistical analysis enabling pragmatic quantification

Psychological 'measurement' (e.g., psychometrics) is aimed at discriminating well and consistently between cases (e.g., individuals, groups) and in ways considered important (e.g., social relevance, relations to future outcomes). Therefore, 'measuring instruments' (e.g., intelligence tests, rating 'scales') are designed

such as to generate data structures that are useful for these purposes (e.g., specific distribution or association patterns). To this pragmatic end, statistical analyses are indispensable (Uher, 2021c).

Many psychologists believe that measurement involves the assignment of numbers and capitalises on their mathematically defined quantitative meaning. In measurement, however, we assign numerical values whose specific *quantitative meaning* is conventionally agreed and traceable to defined reference quantities (e.g., of the International System, SI; BIPM, 2019). We know this from everyday life. The numerical values of '1 kilogram', '2.205 pounds', '35.274 ounces' and '0.1575 stones' differ—but they all indicate the same quantity of weight. These differences originate from once arbitrary decisions on specific quantities that were used as references. Meanwhile, their specific quantitative meaning is conventionally agreed and indicated by the measurement unit (e.g., 'kg', 'lb', 'oz', 'st'). The unit also indicates the specific kind of property measured—'1' 'kilogram' is not '1' 'litre', '1' 'metre' or '1' 'volt'. That is, the measurement unit specifies also a result's *qualitative meaning*, such as whether it is a quantity of weight, volume, length or electric potential.

In psychology, by contrast, 'measurement' values are commonly presented without a unit, thus indicating neither specific qualities (e.g., frequency, intensity or level of agreement) nor specific quantities of them (e.g., *how* often or *how* much of that). Unit-free values—therefore called 'scores'—are meaningless in themselves. It requires statistics to first create quantitative meaning for scores from their distribution patterns and interrelations within specific samples (e.g., differential comparisons within age groups), leading to reference group effects (Uher, 2021c,d, 2022a, 2023a). Hence, psychometric scores constitute *quantifications that are created for specific uses, contexts and pragmatic purposes*, such as for making decisions or projections in applied settings (Barrett, 2003; Dawes et al., 1989; Newfield et al., 2022). This highlights first important differences from genuine measurement.

Specifically, psychometric theories and empirical practices clearly build on a *pragmatic utilitarian framework* that is aimed at producing quantitative results with statistically desirable and practically useful structures. By contrast, traceable relations to empirical interactions with the quantities to be measured (measurands) in individuals and to known reference quantities are neither conceptualised nor empirically implemented. Nevertheless, psychometricians explicitly aim for "measuring the mind" (Borsboom, 2005)—thus, for 'measuring' specific quantitative properties that individuals are assumed to possess. Accordingly, psychometric results (e.g., IQ scores) are interpreted as quantifications of the studied individuals' psychical[1] properties (e.g., intellectual abilities) and used for making decisions about these individuals (e.g., education). Here, psychometricians clearly invoke the *realist framework* underlying physical measurement, ignoring that they have theoretically and empirically established instead only a pragmatic utilitarian framework (Uher, 2021c,d, 2022b, 2023a). This confusion of two incompatible epistemological frameworks entails numerous conceptual and logical errors, as this article will show (Section 3).

But regardless of this, psychometricians' declared aims and result interpretations highlight basic ideas of measurement that are shared by metrologists, physicists and psychologists alike. These ideas can be formulated as two *epistemic criteria* as the most basic common denominators considered across the sciences that characterise an empirical process as one of measurement. Criterion 1 is the *epistemically justified attribution* of the generated quantitative results to the specific properties to be measured (measurands) in the study phenomena and to nothing else. Criterion 2 is the *public interpretability* of the results' quantitative meaning with regard to those measurands (Uher, 2020b, 2021a,b, 2023a). These two criteria are key to distinguish genuine measurement from other processes of quantification (e.g., opinions, judgements, evaluations). Importantly, this is not to classify some approaches as 'superior' or 'inferior'. Rather, a criterion-based approach to define measurement is essential for scrutinising the epistemic fundamentals of a field's pertinent theories and practices. This allowed for identifying, for example, the epistemological inconsistencies inherent to psychometrics (Uher, 2021c,d). A criterion-based approach is also crucial for pinpointing commonalities and differences between sciences.

Concretely, it shows that proposals to 'soften', 'weaken' or 'widen' the definition of measurement for psychology (Eronen, 2024; Finkelstein, 2003; Mari et al., 2015) are epistemically mistaken. Certainly, psychology does not need the high levels of measurement accuracy and precision, as necessary for sciences like physics, chemistry and medicine where errors can lead to the collapse of buildings, chemical explosions or drug overdoses. But changing the definition of a scientific activity as fundamental to empirical science as that of measurement cannot establish its comparability across sciences. Much in contrast, it undermines comparability because it fails to provide guiding principles that specify how analogues of measurement that appropriately consider the study phenomena's peculiarities can be implemented in other sciences. The methodological principles of data generation traceability and numerical traceability, for example, can guide the design of discipline-specific processes that allow for meeting the two epistemic criteria of measurement also in psychology (Uher, 2018a, 2020b, 2022a,b, 2023a). Labelling disparate procedures uniformly as 'measurement' also obscures essential and necessary differences in the theories and practices established in different sciences as well as inevitable limitations. Ultimately, measurement is not just any activity to generate numerical data but involves defined processes that justify the high public trust placed in it (Abran et al., 2012; Porter, 1995).

In everyday life, the *differences between measurement and pragmatic quantification* are obvious. When we buy apples in a shop, we measure their weight. But we do not measure their price. The apples' weight is a quantitative property, which they possess as real physical objects. It is determined through their traceable empirical interaction with a measuring instrument (therefore, we

1 Here, the terms psych*ical* and psych*ological* (from Greek -*logia* for body of knowledge) are distinguished to express the crucial distinction between ontological concepts describing the objects of research themselves (e.g., mental, emotional) and epistemological concepts describing the means for exploring these objects of research (see Section 4.1). This distinction is made in many languages (e.g., French, Italian, Spanish, German, Dutch, Greek, Russian, Norwegian, Swedish, Danish) but not commonly in the English (Lewin, 1936; Uher, 2021b, 2022b, 2023a).

must place the apples on the weighing scale). The specific quantity of weight that we denote as '1 kg' is (nowadays) specified through known reference quantities, which are internationally agreed and thus, universally interpretable. The apples' price, by contrast, is pragmatically quantified for various purposes within a given socio-economic system that go beyond the apples' specific physical properties (e.g., sales, profit). Thus, the price merely indicates an attributed quantitative value—an *attribute*—which therefore changes across contexts and times (e.g., supply, demand and tariffs). The price's specific quantitative meaning, in turn, is derived from its relations to other attributed socio-economic values (e.g., currency, inflation) and can therefore vary in itself as well.

Psychological 'measurement' (e.g., psychometrics) is widely practised and justified for its pragmatic and utilitarian purposes. However, it does not involve genuine measurement as often claimed (therefore here put in inverted commas, as are the psychological terms 'scales' and 'instruments'[2]). Instead, psychological 'measurement' serves other epistemic purposes for which statistics is indispensable. Its focus is on analysing structures in data sets, such as data on persons' test performances or responses to rating 'scales', in order to derive hypothetical quantitative relations, such as levels of "person ability" or item difficulty in Rasch modelling and item response theory. But the specific ways in which the analysed data—as well as the performances and responses encoded in these data—are generated in the first place are still hardly studied (Lundmann and Villadsen, 2016; Rosenbaum and Valsiner, 2011; Toomela, 2008; Uher, 2015c, 2018a,b, 2021a, 2022a, 2023a; Uher and Visalberghi, 2016; Uher et al., 2013b; Wagoner and Valsiner, 2005).

Indeed, rating 'scales', psychology's most widely-used method of *quantitative data generation*, remained largely unchanged since their invention a century ago (Likert, 1932; Thurstone, 1928). This is astounding given that rating data form the basis of much of the empirical evidence used to test scientific hypotheses and theories, to make decisions about individuals in applied settings (Uher, 2018a, 2022b, 2023a) and to evaluate the effectiveness of interventions and trainings (Truijens et al., 2019b).

Hence, there is a *gap* between psychologists' numerical data and statistically modelled quantitative results, on the one side, and the specific entities to be quantified in their actual study phenomena, on the other. Bridging this gap requires measurement.

## 1.3 Metrological frameworks of measurement: Inherent limitations for psychology

Unlike statistics, measurement concerns how the data are generated—thus, the ways in which they are empirically connected both with the unknown quantity to be measured (measurand) in the study phenomena (data generation traceability) and with known reference quantities (numerical traceability). Unbroken documented connection chains determine how the measurement results can be interpreted regarding these measurands qualitatively and quantitatively (epistemic criteria 1 and 2). These two traceability principles underlie the measurement processes established in metrology (Uher, 2020b, 2022a).

Metrology, however, is concerned solely with the measurement of physical properties in non-living nature that feature *invariant* relations. Such properties are *always related to one another in the same ways* (under specified conditions), such as the fundamental relations between electric voltage (V), current (I) and resistance (R). It is this peculiarity that enables their formalisation in *immutable laws* (e.g., Ohm's law) and non-contradictory *mathematical equations* (formulas, e.g., $V = I * R$). Invariant relations can also be codified in *natural constants* (e.g., gravity on Earth, speed of light) and internationally agreed systems of units (e.g., metric, imperial; JCGM100:2008, 2008). Therefore, physical laws and formulas, natural constants and international units of measurement are assumed to be *universally* applicable.

But precisely because of this peculiarity, metrological frameworks cannot be applied or translated to psychological research as directly as metrologists and psychometricians increasingly propose (e.g., Fisher and Pendrill, 2024; Mari et al., 2021). This is because psychology's objects of research feature peculiarities not known from the non-living 'world'. These involve variability, change and novel properties emerging from *complex relations* leading to irreversible development as well as the non-physicality and abstract nature of experience, and others (Hartmann, 1964; Morin, 1992). Moreover, unlike physical sciences and metrology, psychology explores not just objects and relations of specific phenomena (e.g., behaviours) in themselves but also, and in particular, their *individual (subjective) and socio-cultural (inter-subjective)* perception, interpretation, apprehension and appraisal (Wundt, 1896). These complex study phenomena are described in *multi-referential conceptual systems—constructs*. These conceptual systems cannot be studied with physical measuring instruments but require language-based methods instead (Kelly, 1955; Uher, 2022b, 2023b). Language, however, involves complexities that present unparalleled challenges to standardised quantitative inquiry, as this article will demonstrate. To tackle the challenges posed by psychology's complex study phenomena and methods of inquiry, metrology provides neither conceptual nor methodological fundamentals (Uher, 2018a, 2020b, 2022a).

Attempts to directly apply a science's concepts and theories to study phenomena not explored by that science involve challenges that cannot be mastered using the conceptual and methodological fundamentals of just single disciplines. Such *interdisciplinary*[3] approaches underlie the current attempts to directly apply or translate metrological concepts to psychological 'measurement'

2 The terms 'scales' and 'instruments' are put in inverted commas because they do not enable all methodological functions that genuine measuring scales and instruments fulfil, as shown in this article and in Uher (2022a).

3 *Interdisciplinarity* is the synergistic collaboration of several disciplines who work on a specific research objective or problem whose solution is beyond a single discipline's scope. It is aimed at synthesising and integrating perspectives, knowledge, theories and concepts, whereby approaches and methods are transferred between disciplines and integrated through the research topic into their disciplinary work. Often, interdisciplinary projects benefit just one of the disciplines involved but not the others, or at least not immediately (Russell, 2022; Uher, 2024).

and psychometrics (e.g., Fisher and Pendrill, 2024; Mari et al., 2021). But they overlook fundamental ontological, epistemological and methodological differences. Developing epistemically justified research frameworks that are applicable across the sciences in that they are appropriate to the peculiarities of their different objects of research requires scrutinising the basic presuppositions of all the sciences involved. Such elaborations are at the core of transdisciplinarity, which is therefore applied in this article.

## 1.4 Transdisciplinarity: A new way of thinking and scientific inquiry

*Transdisciplinarity* has gained recognition as a new way of thinking about and engaging in scientific inquiry (Montuori, 2008; Nicolescu, 2002, 2008). Unlike all other types of disciplinary collaboration (e.g., cross-, multi- and inter-[4]), transdisciplinarity is aimed at analysing complex systems and complex ("wicked") real-world problems, at developing an understanding of the 'world' in its complexity and at generating unitary intellectual frameworks beyond specific disciplinary perspectives. To enable such explorations, transdisciplinarity[5] not only relies on disciplinary paradigms but also transcends and integrates them. It is aimed at exposing disciplinary boundaries to facilitate the understanding of implicit assumptions, processes of inquiry and resulting knowledge as well as to discover hidden connections between different disciplines and their respective bodies of knowledge. A key focus is on identifying non-obvious differences, particularly in the underlying *ontology* (philosophy and theory of being), *epistemology* (philosophy and theory of knowing) and *methodology* (philosophy and theory of methods, connecting abstract philosophy of science with empirical research). That is, transdisciplinarity explores research questions that can be comprehended only outside of the boundaries of separate disciplines and therefore challenges the entire framework of disciplinary thinking and knowledge organisation (Bernstein, 2015; Gibbs and Beavis, 2020; Piaget, 1972; Pohl, 2011; Uher, 2024).

The present analyses—spanning concepts and approaches from psychology, social sciences, life sciences, physical sciences and metrology—rely on the *Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals* (*TPS Paradigm*;[6] for introductory overviews, see Uher, 2015b, 2018a, pp. 3-8; Uher, 2021b, pp. 219–222; Uher, 2022b, pp. 3–6). This meta-paradigm was already applied, amongst others, to explore the epistemological and methodological fundamentals of data generation methods (Uher, 2018a, 2019, 2021a) and of theories and practices of measurement and pragmatic quantification across the sciences (Uher, 2020b, 2022a) as well as to scrutinise those underlying psychometrics and quantitative psychology (Uher, 2021c,d, 2022b, 2023a). Pertinent key problems were demonstrated empirically in multi-method comparisons (e.g., Uher et al., 2013a; Uher and Visalberghi, 2016; Uher et al., 2013b). The present article builds upon and substantially extends these previous analyses.

## 1.5 Outline of this article

This article offers a novel and ambitious transdisciplinary approach to advance the epistemological and methodological fundamentals of quantitative psychology by integrating relevant concepts from mathematical biophysics, metrology, linguistics, complexity science, psychology and philosophy of science. It elaborates the epistemic process structure of measurement, highlighting crucial differences to statistics (e.g., psychometrics). A focus is on elaborating the ways in which the peculiarities of

4 Coss-disciplinary, multi-disciplinary or inter-disciplinary collaborations are sometimes erroneously referred to as 'transdisciplinary', ignoring the fundamental differences between them (Bernstein, 2015; Uher, 2024).

5 There are two schools of transdisciplinarity. The present analyses build on *theoretical transdisciplinarity*. *Applied (practical) transdisciplinarity*, by contrast, is aimed less at developing theoretical frameworks and new forms of knowledge but more at understanding complex real-world problems and developing tangible solutions. It involves scholars from different disciplines but also political, social and economic actors as well as ordinary citizens with the aim of producing socially robust knowledge rather than merely reliable scientific knowledge (Uher, 2024).

6 The TPS Paradigm is aimed at making explicit the basic presuppositions that different disciplines (e.g., biology, medicine, psychology, social sciences, physical sciences) make about research on individuals and their multi-layered 'realities' considering phenomena from all domains of human life. These involve abiotic phenomena (e.g., non-living environment), biotic phenomena (e.g., physiology, behaviours), psychical phenomena (e.g., emotions, thoughts) and socio-cultural phenomena (e.g., culture, language), which are all merged in the single individual and its functioning and development but involve different layers of 'reality'. This poses challenges for empirical inquiry because different kinds of phenomena require different ontologies, epistemologies, theories, methodologies and methods, which are based on different, even contradictory basic presuppositions (Uher, 2024).

To provide conceptual fundamentals that are appropriate to tackle these challenges and to discover hidden connections between scientific disciplines and their knowledges, relevant established concepts from various sciences have been systematically integrated on the basis of their basic presuppositions and underlying rationales and complemented by novel ones. This enabled the development of three unitary frameworks that coherently build upon each other (therefore, it is termed a 'paradigm'), that transcend disciplinary boundaries (therefore termed 'transdisciplinary'), and that are aimed at making explicit the most basic assumptions made in a field (therefore termed philosophy-of-science). The *philosophical framework* comprises presuppositions for research on individuals (e.g., complexity, complementarity, anthropogenicity). The *metatheoretical framework* comprises, amongst others, metatheoretical definitions of various kinds of phenomena studied in individuals, differentiated by their modes of accessibility to humans (e.g., physiology, psyche, behaviour, sign systems like language; Uher, 2013, 2015b,c, 2016a,b, 2023b). It informs the *methodological framework*, which comprises, amongst others, classifications of data generation methods based on the modes of accessibility that they enable; basic principles, concepts and theories of measurement and quantification across the sciences demonstrated in empirical multi-method comparisons as well as critical analyses of the foundations of psychometrics and quantitative psychology: http://researchonindividuals.org.

language, when used in psychological methods (e.g., rating 'scales', variables and models), obscure the epistemic differences between them. This confusion contributes to the common yet erroneous belief that statistics could constitute psychology's approach for 'measuring' its study phenomena. The analyses are made with regard to psychology but equally apply to pertinent practices in other sciences.

Section 2 introduces fundamentals of measurement. These involve the measurement problem—the epistemically necessary distinction between the object of research and the objects used as measuring instruments as well as the conceptualisation of how the latter can provide information on the former. Measurement also requires the formal representation of observations in sign systems (e.g., data, formal models). The section presents Rosen's system of modelling relations as an abstract general model of the entire measurement process—from (1) conceptualising the objects of research, over (2) generating the data, (3) formally manipulating these data (e.g., statistical analysis) up to (4) interpreting the formal outcomes obtained with regard to the actual study phenomena. This process model is shown to underlie metrologists' approaches for tackling the problem of circularity in physical measurement, illustrated in the special cases of measurement coordination and calibration.

Section 3 applies these fundamentals to explore the challenges involved in establishing genuine analogues of measurement in psychology, which arise from the peculiarities of its study phenomena (e.g., higher-order complexity, non-ergodicity) and those of the language-based methods required for their exploration (e.g., inbuilt semantics). It demonstrates that psychology's focus on statistical modelling (e.g., psychometrics)—thus, on just one of the four necessary and interrelated modelling relations in Rosen's scheme—ignores the entire measurement process. But this often goes unnoticed because researchers consider only the general (dictionary) meanings of their verbal 'scales'—their inbuilt semantics, yet ignore how raters actually interpret and use these 'scales'. This introduces several breaks in the data's and model's relations to the actual phenomena that these are meant to represent. It also obscures psychology's measurement problem. This involves not just the crucial distinction between the phenomena studied (e.g., feelings) and those used as 'instruments' for studying them (e.g., descriptions of feelings) but also individuals' (e.g., raters') local context-specific interactions with both. These complexify the ways in which epistemically justified (valid) information about the study phenomena can be obtained through language-based methods.

Section 4 shows that the frequent failure to distinguish the study phenomena from the means of their investigation (e.g., 'instruments', formal models) confuses ontological with epistemological concepts—psychologists' cardinal error. This logical error is fuelled by quantitative psychologists' focus on statistics as well as by our human tendency to mistake verbal descriptions for the phenomena described. Many psychologists therefore mistake judgements of verbal statements for measurements of the phenomena described. Many also overlook that statistics can neither establish nor analyse a formal model's relations to the real phenomena studied. Establishing these relations requires genuine analogues of measurement for which the section elaborates necessary epistemological and methodological fundamentals. It closes by showing ways in which the powerful artificial intelligence systems (AI) now available for modelling human language can meaningfully support psychological research but also perpetuate psychologists' cardinal error.

## 2 Key problems of measurement

Measurement, in its most general sense, is a highly selective form of observation because 'to measure' means that we must choose to measure *something* without having to measure *everything*. Every object of research may feature various non-equivalent properties (e.g., length, temperature and weight) as well as different quantitative entities of the same property (e.g., foot length, finger length and body height). Measurement is a process that involves the detection and recognition of selected properties in the object researched and that produces justified information about them (von Neumann, 1955; Uher, 2022b).

For simplicity, when 'objects' are mentioned in the following, this is always meant to include their properties as well because we cannot measure objects in themselves (e.g., physical bodies) but only their specific properties (e.g., mass, voltage and temperature). Properties are also included when we understand by 'objects of research' not just physical objects (e.g., individuals' bodies) but also non-physical phenomena (e.g., individuals' reasoning, beliefs and emotions)—thus, denoting the subject matter in general.

### 2.1 The measurement problem: Distinguishing the objects of research from the objects used as measuring instruments and conceptualising their interaction

We can describe all objects in their existence and being in the 'world', thus ontologically. To describe how we can gain knowledge about a given object, thus epistemologically, we must distinguish the ontic object (the specific concrete entity) to be measured from the objects used for epistemic (knowledge-generating) purposes as measuring instruments. *Measurement* defines a theory-laden process structure that conceptualises the objects of research and the methods (including instruments) used to gain epistemically justified information about them (von Neumann, 1955).

Specifically, measurement requires an *empirical interaction* between the specific quantity to be measured (measurand) in the study object (e.g., the temperature of a cup of coffee) and the object used as instrument (e.g., mercury in glass tube). Measuring instruments must be designed such that they produce, through their empirical interactions with the measurand, distinctive *indications* that are observable for humans. In iterative processes of theorising and experimentation, scientists identify which variations of an instrument—when applied in defined ways (the *method*)—reliably produce distinct and for humans easily discernible patterns (e.g., linear extension of mercury in glass tubes). These indications are used to make inferences on the study object's specific state at the moment of interaction to obtain *information* about it. That is, scientists use their current state of knowledge to *decide* how to design specific objects as instruments, how to use them (methods) and which indications of their empirical interactions with the

study object to consider as informative—thus, how 'to read' these indications (Mari et al., 2021; Pattee, 2013; Tal, 2020; Uher, 2020b, 2023a).

In sum, the *measurement problem*[7] concerns the epistemic distinction of the object of research from the objects used as measuring instrument. It requires their conceptualisation as well as that of their presumed empirical interaction under defined conditions (method) producing observable indications. To document and analyse them to derive measurement results, the observed and interpreted indications must be formally represented.

## 2.2 Measurement requires semiotic representation in rule-based formal models

The relations between physical properties are empirically given, invariant and lawful (those studied in metrology). But information about them can be *formalised* in various ways. Formalisms are conceptual, mathematical, algorithmic, representational and other abstract operations that follow logical, deductive or arbitrarily prescribed *rules*. In measurement, *formal representation* involves sign systems. Signs are composed of tokens (sign carriers; e.g., Latin or Greek letters, Arabic numerals) that are assigned meanings, which specify the information that these tokens are meant to represent (e.g., specific indications observed or quantitative relations). These sign systems constitute the *data* and formal models (e.g., variables, numerical values), which can be used to analyse the information represented (e.g., mathematically). The signs' meanings, however, because they just are ass*ign*ed (attributed and ascribed), can vary. Numerals can represent numbers but also just order (e.g., door 'numbers') or just nominal categories (e.g., genders). That is, formalisation is *arbitrary, non-physical and rule-based* (Abel, 2012; Pattee, 2013; Uher, 2023a; von Neumann, 1955).

In sum, semiotic (sign-based) representation is essential for all empirical sciences (Frigg and Nguyen, 2021; Pattee, 2013; van Fraassen, 2008). It requires that data and models are clearly distinguished from the objects that they semiotically represent. This separation is no philosophical doctrine but an epistemic necessity that follows from the *definition of a sign as something that stands for something other than itself* (Pattee, 2001; Peirce, 1958; Uher, 2020b, 2022b). The ways in which interpreted observations are encoded into data in a study are therefore crucial for understanding and analysing these data. The specific encoding is also essential for drawing justified conclusions from the analytical results about the actual objects explored. The study objects, their formal representations and the interrelations between both can be conceptualised and analysed in an overarching model.

## 2.3 The system of interrelated modelling relations underlying empirical science

Robert Rosen, a mathematical biophysicist and theoretical biologist, developed a general relational model to conceptualise the processes by which living beings selectively perceive specific parts of their environment and make sense of that information. Scientific knowledge generation is a special case of these fundamental processes. Rosen (1985, 1991, 1999) developed this process model mathematically building on earlier work by Rashevsky (1960b,a) and using category theory (Lennox, 2024).

### 2.3.1 Category theory: Modelling the relations of relations between objects

Many psychologists associate mathematics solely with quantitative analysis (e.g., algebra, arithmetic, calculus). But mathematics also involves many non-quantitative branches, such as category theory, combinatorics, geometry, logic, set theory or topology (Linkov, 2024; Rudolph, 2013), which are also used in empirical sciences.

*Category theory* is a general mathematical theory to formally describe abstract structures and relations. In this theory, a category is a system of mathematical objects and their relations. The focus is on conceptualising these *relations, understood as morphisms, arrows or functors*, that map a source object to its target object in specific ways (e.g., through structure-preserving transformations). Category theory also permits to map these relations in themselves—thus, to map the relations between categories, termed *natural transformations*. Hence, category theory is about modelling (mathematical) objects, relations of objects as well as relations of relations (Leinster, 2014). This makes it suitable to model also the process of scientific modelling in itself (Rosen, 1985).

### 2.3.2 Scientific modelling: Modelling the relations between causality, encoding, analysis and decoding

For scientific inquiry in general, Rosen's system[8] of interrelated modelling relations conceptualises the basic set of processes that are used to explore a specific part of the 'world', conceived as the

7 This is one of the most intricate and also variously defined problems of quantum physics. It arose because its micro-physical objects of research (e.g., electrons) cannot be made accessible to observers other than through their interactions with macro-physical instruments (e.g., detection screen). This entails challenges called the measurement problem. Simply put, quantum physicists sought to explain how the macro-physical instruments can provide information about micro-physical objects, thus what constitutes a measurement. This required the conceptual distinction between the object of research and its environment (incl. the instruments and methods of observation)—the *Heisenberg cut*. It also required explanation of the processes by which the micro-physical objects under study interact with the macro-physical objects used as measuring instruments, which however are debated still today (Atmanspacher, 1997; Hance and Hossenfelder, 2022; Heisenberg, 1927; von Neumann, 1955).

8 Rosen himself (Rosen, 1985, 1999) and others refer to this process model solely as *modelling relation*. To highlight that it involves the coherent modelling of four interrelated modelling relations (arrows 1 to 4) and to pinpoint key distinctions to the statistical modelling of data, which concerns solely arrow 3 in Rosen's general model, I refer to his process model as a *system of interrelated modelling relations*.

real system under study (object of research, study phenomena). These processes specify the ways in which this *real system being studied is mapped to the formal system that is used for studying it.* Stated in category-theoretic terms, these modelling relations relate disjoint categories of objects (Mikulecky, 2000, 2001). In everyday life, we intuitively establish such modelling relations whenever we try to make sense of the complex phenomena that we encounter, grounded in the general belief that these are not completely random but show some kind of order. Figure 1 illustrates the system of interrelated modelling relations, comprising the real study system and the formal system used for studying it as mathematical objects as well as the processes (mappings, relations) that are conceptualised within and between them, depicted as arrows. What do these different processes involve?

In science (and everyday life), when we perceive events as changes (e.g., in behaviour), we attribute to those changes some causes that we seek to explain (e.g., mental abilities, intentions) as possible causes of the observed events (e.g., through abduction; Peirce, 1958, CP 7.218). This (presumed) causal relation in the real system (e.g., a person) is depicted as arrow 1. Its exploration requires the *encoding* of the real changes observed. That is, selected indications that we deem relevant for exploring the presumed causal relations are encoded into objects and relations in the formal study system. These encoding relations[9]—the *data generation*—are depicted as arrow 2. The formal system is the explicit scientific model (or, in everyday life, the intuitive mental model) that we create to deal with the information obtained from our selected observations. It serves as a *surrogate* system that we can explore in ways that are not possible with the real system itself, such as mathematical analysis rather than physical dissection. Hence, the model is analysed *in lieu of* the actual objects of research (Rosen, 1985, 1991; Uher, 2015b,c,d).

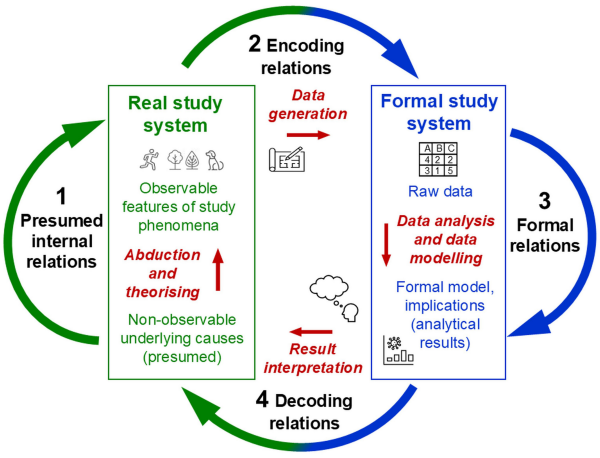



FIGURE 1
Rosen's general process structure of empirical science: A coherent system of four interrelated modelling relations. The real study system and the formal system used for studying it, conceptualised as mathematical objects, as well as the processes (mappings, relations) each within and back and forth between them, depicted as arrows. Adapted from Rosen (1985) and Uher (2022b, Figure 6).

We can manipulate the information encoded in the formal system in various ways using data modelling techniques (e.g., statistical or algorithmic analysis) to try to imitate the causal events that presumably occur in the real system (e.g., simulation models). Therefore, we must use our current knowledge of that real system (e.g., a person), its observable indications (e.g., behavioural responses) and (possible) non-observable internal relations (e.g., mental abilities) to decide which specific operational manipulations (e.g., statistical analysis) are appropriate to explore the information about that real system. Through manipulative changes and operations performed in the formal system—the *data analysis*—depicted as arrow 3, we obtain an *implication*, such as statistical or simulation results.

Once we believe that our formal system (e.g., structural equation model) is appropriate and may correspond to the presumed causal events in the real system, we must relate the results obtained in the formal system back to the real system studied. This *decoding* relation, depicted as arrow 4, requires interpreting the *formal* results with regard to the *non-formal* events occurring in the real study system. The aim is to check how well the formal model may represent the causes that we presume and that could explain the changes observed in that real system. Thus, decoding involves a mapping relation between disjoint categories of objects—thus, between the outcomes generated in a *formal* study system (e.g., mathematical) and the outcomes observable in a *real* study system (e.g., behavioural).

If the processes of encoding (2), implication (3) and decoding (4) appears to reproduce the presumed causal processes (1) sufficiently accurately, the system of modelling relations it said to *commute*. Commutation implies that the formal study system established in this process constitutes a successful model of the real system studied—expressed in category-theoretic terms by the equation: $1 = 2 + 3 + 4$. Note that these numerals represent not numbers but different kinds of mapping relations, depicted as the four arrows in Figure 1. Hence, the system of modelling relations *conceptualises the relations between relations between objects* of different kinds (Rosen, 1985, 1991, 1999).

Rosen's process model is not commonly taught. Many scientists are even puzzled when they first encounter it (Mikulecky, 2011). This is astonishing and unfortunate because it conceptualises how empirical science, in general, and measurement, in particular, are done.

### 2.3.3 How empirical science is done: The epistemic necessity of making subjective decisions

Rosen's system of interrelated modelling relations highlights several key points that are fundamental to empirical inquiry but often not well considered. First, it specifies that the system studied and the surrogate system (model) used for studying it are of different kinds—*real vs. formal*. The relations (mappings) established between them—encoding (arrow 2 in Figure 1) and decoding (arrow 4)—therefore involve transformations that cannot

9 Also called *rules of correspondence* (Kaplan, 1964/2017; Margenau, 1950; Torgerson, 1958).

be derived from within either system. These relations are thus independent of both systems.

Specifically, potentially unlimited amounts of observations that can be made of a real study system must be mapped onto the limited sign system that is used as its formal model. Encoding therefore requires that scientists reduce and simplify their observations to only those elements that they interpret as relevant for their given research question and that they choose to encode as data. Thus, the essence of encoding is *high selectivity* and *reduction*. This requires *representational decisions* about what to represent, and what not, and about how to represent it (Harvard and Winsberg, 2022). For example, observations of variable and highly dynamic phenomena, such as behaviours (e.g., hand gestures), often require their encoding in fuzzy categories. This involves the mapping of fuzzy subsets of observations (e.g., physical states of fingers) into the same formal category (e.g., hand configurations; Allevard et al., 2005). That is, scientific representation, in general, and measurement, in particular, involves the *selective reductive mapping of an open domain of a study system to a closed sign system used as its surrogate model* (for general principles, see Uher, 2019).

Decoding—the inverse relation from the formal system back to the real system (arrow 4)—as well, is a delicate process that is prone to many potential points of failure. This is because it involves the transformation of results obtained through *formal* manipulations (e.g., mathematical, statistical), which are not possible in the *real* system (e.g., behavioural, psychical) itself (Mikulecky, 2000; Rosen, 1985, 1999). This epistemic necessity makes the modelling process prone to *methodomorphism*, whereby methods impose structures onto the results that, if erroneously attributed to the study phenomena, may (unintentionally) influence and limit the concepts and theories developed about them (Danziger, 1985; Uher, 2022b).

Second, Rosen's process model highlights that the only part of our scientific models that—taken by itself—is free from operational subjectivity is the formal study system (e.g., statistical model) that is used as a surrogate for the real system studied (arrow 3). However, the formal model is established by the scientists' choice and decisions and is therefore subjective in many ways as well (Mikulecky, 2000, 2011; Rosen, 1991; Strauch, 1976).

> "This makes modelling as much an *art* as it is a part of science. Unfortunately, this is probably one of the least well appreciated aspects of the manner in which science is actually practised and, therefore, one which is often actively denied" (Mikulecky, 2000, p. 421).

In sum, Rosen's general model conceptualises the processes of empirical science that epistemically justify the representation of observable regularities by means of abstract (e.g., mathematical) models. These processes concern the *coordination (or correspondence) between theory and observable phenomena*, such as the applicability of theoretical concepts to concrete events—known as the *problem of coordination (or correspondence)* in science (Hempel, 1952; Margenau, 1950; Torgerson, 1958). To specify the conditions under which abstract representations can be applied to observable phenomena and used to investigate—and also to quantify—entities of non-observable phenomena, it requires measurement.

## 2.4 Tackling the epistemic circularity of measurement requires a coherent system of modelling relations

Any method of data generation involves categorisation, which enables basic forms of analysis, such as grouping or classifying objects by their similarities and differences. Measurement has advantages over mere categorisation[10] by enabling more sophisticated analyses of categorised objects and their relations by additionally enabling the descriptive differentiation *between instances that are of the same kind (quality) and divisible—thus, that differ in quantity* (see Hartmann, 1964; Uher, 2018a, 2020b).

Key problems of measurement arise from the fact that many objects of research are not directly observable with our senses (e.g., electric potential, others' mental processes) or not accurately enough (e.g., weight of smaller objects). Rosen's process model underlies the approaches that are used to tackle these epistemic challenges, as illustrated here in the problems of measurement coordination and calibration.

### 2.4.1 Measurement coordination: Exploring the relations between observable indications and unobservable measurands

*Measurement coordination* is the specific problem of how to justify the assumption that a specific measurement procedure does indeed allow us to measure a specific property in the absence of independent methods for measuring it. This involves the problem of how to justify that specific quantity values are assigned to specific measurands under a specific methodical procedure. Measurement coordination (also "problem of nomic measurement"; Chang, 2004) thus concerns the relations between the abstract terms used to express information about quantities and the ways of measuring those quantities (Luchetti, 2020).

Challenges arise from many phenomena's non-observability. We can often directly observe neither the specific quantity to be measured (measurand; e.g., a body's temperature) nor its relation to the observable quantitative indications that are produced by its interaction with the measuring instrument (e.g., length of mercury in glass tubes) and that may be useful to infer the measurand's unknown quantity. Thus, in the early stages of scientific inquiry, the mapping relation between indications and measurands is unknown (e.g., the function relating the values of length of mercury with temperature). But it cannot be determined empirically without already established, independent measurement methods—because it is through measurement that such relations are first established. This requires scientists to make preliminary decisions about what counts as an indication of the property studied (e.g., temperature)—not knowing their specific relations, nor (initially) what exactly that property actually is, nor what other factors may influence an instrument's observable indications.

The fact that these questions cannot be addressed independently of each other involves *epistemic circularity*,

10 Therefore, mere nominal categorisation should not be confused with measurement.

discussed in many sciences and philosophy for a century already (Chang, 2004; Luchetti, 2024; Mach, 1986; Reichenbach, 1920; van Fraassen, 2008). To tackle this problem, scientists must establish appropriate and independent sources of justification for a specific measurement procedure and the assignment of specific values to specific quantities of a specific property. To achieve this, they must coordinate several modelling relations and establish their interrelations coherently.

To construct thermometers, for example, scientists began with preliminary definitions that coordinated a preliminary theoretical concept of temperature with empirical indications that could be obtained from preliminary instruments and their variations. They filled various liquids or gases in glass tubes and studied variations in their extension (volume) obtained from various heat-producing operations. Presuming a linear invariant relation between volume and temperature, scientists experimented with different substances (e.g., alcohol, hydrogen, mercury and water) to identify under which standardised conditions (e.g., pressure and heat production) which substance reliably produces distinct (e.g., monotonously increasing) indications, thus showing thermometric properties. From consistent indications produced by different thermometric substances, scientists could develop different kinds of thermometers, thus enabling triangulation. The redefinition of temperature as the average kinetic energy of particles provided a theoretical foundation to substantiate the linear invariant relation between temperature and the volume of specific substances used in thermometers (under specified conditions; Chang, 2004; JCGM100:2008, 2008; Kellen et al., 2021; Uher, 2020b).

The problem of measurement coordination and its inevitable epistemic circularity can thus be tackled through iterative processes in which a coherent system of assumptions is established to justify specific knowledge claims—using a *coherentist approach* (Olsson, 2023). With each epistemic iteration, the theoretical concept is re-coordinated to more reliable indications, which in turn enables more precise tests of predictions, more advanced theories, more refined and more standardised methods and instruments of measurement, and so on (Luchetti, 2024; Tal, 2020; van Fraassen, 2008). Through these *iterative feedback loops*, scientists systematically develop epistemic justifications for having implemented *coordinated connections* between the (presumed) non-observable measurand (e.g., a cup of coffee's specific temperature), the observable indications produced by its lawful (invariant) interaction with the measuring instrument (e.g., length of mercury), a known reference quantity (e.g., another thermometer used for calibration), and the semiotic representations of the information thus-obtained (e.g., '37°Celsius', '98.6° Fahrenheit'). This information is then mathematically analysed in the formal system. The obtained result can be used to make justified inferences on the specific quantity of the non-observable measurand.

Rosen's general model allows for conceptualising the process structure underlying measurement coordination. Accordingly, this involves modelling the presumed relations within the real study system, comprising the non-observable object of research (measurand), the object used as instruments and the observable indication produced from their (non-observable) empirical interaction. Their presumed causal relations (arrow 1 in Figure 2) are then explored empirically through *unbroken*

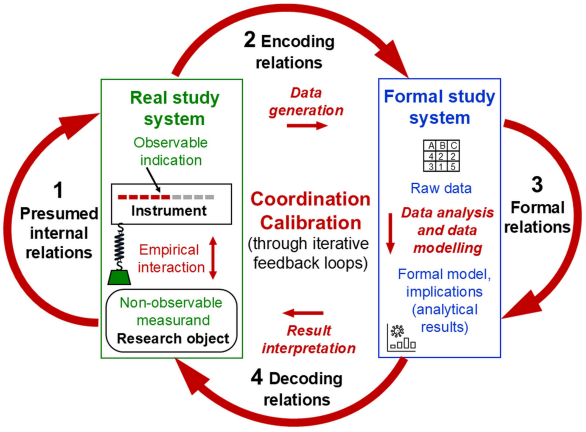



FIGURE 2
Physical measurement: A coherent system of four coordinated and calibrated modelling relations.

*documented traceable* relations to, within and back from the formal system that is used to study that real system (arrows 2, 3 and 4). In iterative feedback loops, the four modelling relations in Rosen's system (arrows 1 to 4) are passed through over and over again, thereby *re-coordinating* them with one another until their commutativity is established, indicating successful modelling of the real study system.

Necessarily, scientists can start to establish measurement coordination only from preliminary assumptions and theories about the study property and from preliminary instruments, methods and decisions on arbitrary encoding rules to obtain first empirical data. They must use preliminary, yet theoretically informed, analytical operations to obtain possibly informative implications. When decoding and interpreting these analytical results, scientists can also make only preliminary assumptions about the implications that these may have for the presumed relations between instrument indications and measurand. Each iteration in the overarching model of a measurement process enables new theoretical, methodical and empirical insights and refinements, which mutually stimulate each other, leading to cascades of development through which a coherent system of epistemically justified knowledge claims is established.

These iterative processes also involve testing and adjusting the specific parameters of a given measurement procedure—through calibration.

### 2.4.2 Calibration: Modelling precision and uncertainty in measurement

*Calibration* procedures establish reliable relations between the instrument indications obtained under a given method in the real study system and the measurement results obtained in the formal model, which specify information about the actual (non-observable) quantity to be measured (measurand). Calibration is theoretically constructed and empirically tested by modelling uncertainties and systematic errors under idealised theoretical and statistical assumptions (e.g., about distribution patterns and the

randomness of influencing factors). The aim is to improve the accuracy of the measurement results by specifying the ranges of uncertainties and errors for all parameters involved in a given measurement procedure. This allows for incorporating corrections for systematic effects (e.g., of pressure on temperature) and for adjusting inconsistent observations of instrument indications (Chang, 2004; Luchetti, 2020; McClimans et al., 2017; Tal, 2017).

That is, calibration involves modelling activities that are aimed at refining the coordinated structure of a measurement process. In Rosen's scheme, this means that the parameters used to establish proportional (quantitative) relations in the measurement model are adjusted within and across all four modelling relations (arrows 1 to 4 in Figure 2). These modelling relations are passed through in iterative feedback loops to obtain quantitative parameter value ranges that maximise the predictive accuracy of the overarching model. Thus, *calibration* refers to the coordination of abstract quantity terms in the formal model with the specific quantities to be measured in real study objects when a specific measurement method (including measuring instrument) is used (Luchetti, 2020; McClimans et al., 2017).

This model-based view of calibration illustrates the coherentist approach that is necessary to tackle the epistemic circularity of measurement. This involves establishing theoretical and empirical justifications for the assumption that a specific method (including instrument) enables the measurement of a specific property in absence of other independent methods for measuring it. Once different methods (and instruments) for measuring the same property (e.g., temperature) are developed, uncertainties and systematic errors can also be modelled across different procedures and instruments, such as to calibrate thermometers involving different kinds of thermometric substances (e.g., gases and fluids; Chang, 2004).

Calibration processes are necessary to implement numerical traceability[11]—thus, to establish for the numerical values used as measurement results a publicly interpretable meaning regarding the specific quantities measured (*how* much of the studied property that is; Uher, 2022a). To ensure that measurement results are reliably interpretable and represent the same quantitative information regarding the measurands across time and contexts (e.g., specific weight of 1 kilogram), metrologists defined primary references, which are internationally accepted (e.g., through legislation) and assumed to be stable (e.g., prototype kilogramme[12]). From each primary reference, large networks of unbroken documented connection chains were established (via national references) to all working references that are used in measurement procedures in research and everyday life (e.g., laboratory weighing scales, household thermometers; JCGM200:2012, 2012). These *calibration chains* specify uncertainties and errors as quantitative indications of the quality of a measurement result to assess its precision and accuracy (JCGM100:2008, 2008; Uher, 2020b).

### 2.4.3 The theoretical and empirical process structure of measurement: A coordinated and calibrated system of four interrelated modelling relations

The essence of measurement is thus a theory-laden process structure that involves modelling relations each within a real and a formal study system as well as back and forth between them, which are coherently connected with one another in an overarching process, as conceptualised in Rosen's general model (Figure 2). This requires data generation methods that enable empirical interactions of the non-observable quantities to be measured with a measuring instrument. Identifying observable indications of these interactions that are (possibly) informative about these measurands requires a general model of coherent and epistemically justified interrelations within and between the real and the formal study system. These are *re*-coordinated and *re*-calibrated with one another by empirically *re*-testing the presumed relations (e.g., comparing predicted and observed indications), *re*-adjusting their parameters (e.g., errors, uncertainties) and *re*-fining assumptions (e.g., randomness).

In sum, a coordinated and calibrated system of interrelated modelling activities is necessary to empirically implement unbroken traceable connection chains that establish proportional (quantitative) relations between the measurement results obtained in the formal model and both (1) the measurand's unknown quantity (data generation traceability) and (2) a known reference quantity (numerical traceability) in the real study system. Measurement models thus-developed allow us to derive from defined observable instrument indications calibrated measurement results that can be (1) justifiably attributed to the measurands, and (2) publicly interpreted in their quantitative meaning regarding those measurands—the two epistemic criteria of measurement across sciences (Uher, 2020b, 2022a). The insights gained from iteratively developing the process structure of a measurement model may also necessitate a revision of the definitions and theoretical explanations of the objects and relations in the real system (e.g., temperature redefined as average kinetic particle energy).

Clearly, physical measurement procedures cannot be directly applied to psychology. But what specifically are the challenges for devising *analogous* processes in psychology?

## 3 Psychology's inherent challenges for quantitative research

The history of metrology testifies to the challenges involved in tackling the problems of measurement coordination and calibration in physical measurement (Chang, 2004)—thus, in the study of invariant relations in non-living nature, which can therefore be formalised in immutable laws, natural constants and mathematical formulas. Psychology, however, explores phenomena (e.g., behaviours, thoughts and beliefs) that are—*in themselves*—variable, context-dependent, changing and developing over time (Uher, 2021b). Such peculiarities are characteristic of living systems

11 Numerical traceability is the transdisciplinary term to denote—on more abstract levels of consideration—the basic principle underlying the concept of metrological traceability used in physical measurement (JCGM200:2012, 2012) in order to adapt it to the peculiarities of non-physical research.

12 The standard unit of one kilogram, originally specified through artefacts, was recently defined in terms of natural constants using the Planck constant, speed of light and the Caesium atom's resonant frequency (BIPM, 2019).

(e.g., psyche and society) and not studied in metrology. These peculiarities entail that the low replicability of psychological findings is not just an epistemic problem that could be remedied with more transparent and robust methods, as many currently believe. Rather, it is also a reflection of the indeterminate variability and changeability of the study phenomena themselves (arrow 1 in Figure 1). Low replicability of psychological findings thus reflects not just epistemic uncertainty of 'measurement' but also *fundamental ontic indetermination* (Scholz, 2024).

## 3.1 Psychology's study phenomena: Peculiarities of higher-order complexity

Living systems (e.g., biotic, psychical and social) are of higher order complexity. They feature peculiarities not known from non-living systems (Baianu and Poli, 2011; Morin, 2008).

### 3.1.1 Emergent properties not present in the processes from which they arise

In higher-order (super) complex systems, interactions occur between various kinds of processes on different levels of organisation from which novel properties emerge on the level of their whole that are not present in the single processes from which they arise. These novel, higher-level properties can also feed back to and change the lower-level processes from which they emerge. Such dynamic multi-level feedback loops lead to continuous change and irreversible development on all levels of organisation (Morin, 1992; Rosen, 1970, 1999).

Human languages, for example, gradually emerged from individuals' interactions with one another. The language of a community, in turn, mediates and shapes the ways in which its single individuals perceive, think and organise their experiences into abstract categories. Through dynamic multi-level feedback processes over time, individuals, their community and their language mutually influence each other, thereby developing continuously further and getting ever more complex (Boroditsky, 2018; Deutscher, 2006; Valsiner, 2007; Vygotsky, 1962). This entanglement of mind and language first enables the use of language-based methods in science. But the intricacies of language also promote conceptual confusions, which are still largely overlooked, as this section will show.

Emergence also entails complex relations between the levels of parts and wholes.

### 3.1.2 Complex wholes and their parts: One–to–many, many–to–one and many–to–many relations

In living systems (e.g., individuals), the same process (e.g., a specific feeling) can generate different outcomes (e.g., different behaviours) in different times, contexts or individuals—thus, involving *one–to–many relations* (multifinality, pluripotency). Vice versa, different processes (e.g., of abstract thinking) can generate the same outcome (e.g., solving the same task)—thus, involving *many–to–one relations* (equifinality, degeneracy; Cicchetti and Rogosch, 1996; Mason, 2010; Richters, 2021; Sato et al., 2009; Toomela, 2008; Uher, 2022b). To consider multiple processes and outcomes at once, we must conceptualise *many–to–many relations* between the parts and their whole on different levels of organisation.

This entails that specific relations from observable indications to non-observable phenomena that apply *to all individuals in all contexts and all times* cannot be identified. This complicates the possibilities for solving the problem of *measurement coordination in psychology*. Specifically, complex relations challenge the appropriateness of the sample-level statistics commonly used in psychology, which are aimed at identifying invariant[13] (e.g., cause–effect) relations, such as between latent and manifest variables in factor analyses or structural equation models—that is, one–to–one relations.

Complex multi-level relations also entail the fact that the properties of parts identified in isolation (e.g., cells) cannot explain the whole (e.g., organism) because its properties emerge only from the parts' *joint* interactions. Changes in single parts or single relations between them can change the properties of the whole. Psychical processes cannot even be isolated from one another, although they can be qualitatively distinguished (Luria, 1966). Thus, *complex wholes are more than and different from the sum of their parts* (Morin, 1992, 2008; Nowotny, 2005; Ramage and Shipp, 2020; Uher, 2024). All this entails that living systems cannot be explored by reducing them to the parts of which they are composed (e.g., organisms to cells), as this is possible for the non-living systems (e.g., technical) featuring invariant relations as studied in metrology (Rosen, 1985, 1991).

### 3.1.3 Humans are thinking intentional agents who make sense of their 'world'

Psychologists also cannot ignore the fact that humans are thinking agents who have aims, goals and values that they pursue with intention and who can anticipate (mentally model) future outcomes and proactively adjust their actions accordingly. Humans hold personal (subjective) and socio-cultural views on their 'world', including on the psychological studies in which they partake. Individuals memorise and learn. Therefore, simple repetitions of identical study conditions (e.g., experiments and items) cannot be used (Danziger, 1990; Kelly, 1955; Shweder, 1977; Smedslund, 2016a; Uher, 2015a; Valsiner, 1998).

In sum, psychology's study phenomena feature peculiarities that do not occur in the properties amenable to physical measurement. These peculiarities complicate the design of analogous research processes that meet the two epistemic criteria of measurement. In the following, we explore these complications stepwise, starting with the level of data analyses.

13 Invariance here refers to *what kinds of objects* are always related to one another (one–to–one rather than e.g., one–to–many), not how. Their specific relations may have quantitatively different forms (e.g., linear, non-linear).

## 3.2 Psychology's focus on aggregate level analysis

Psychology's primary scientific focus (unlike sociology's) is on the *individual*, which constitutes its *theoretical unit* of analysis. The *empirical units* of analysis in psychological 'measurement', however, are *groups*. Why is that so? And what justifies the assumption that results obtained on aggregate levels are suited to quantify individual level phenomena?

### 3.2.1 Indefinitely complex and uncontrollable influence factors: Randomisation and large sample analyses

Unlike metrologists and physical scientists, psychologists cannot isolate their study objects and experimentally manipulate the (presumed) quantities to be measured in them, such as individuals' processing speed, reasoning abilities or beliefs (Trendler, 2009). Moreover, in physical measurement, influencing factors involve comparably few and exclusively here-and-now factors. By contrast, the factors influencing psychology's study phenomena, such as internal and external conditions causing mental distraction, are indefinitely complex and ever-changing and can even transcend the here-and-now (Barrett et al., 2010; Smedslund et al., 2022; Uher, 2016a).

To deal with these challenges, psychologists study groups of individuals that are assumed to be sampled randomly with regard to these unspecifiable and uncontrollable influence factors. To estimate the impact of these factors, psychologists analyse samples that are large enough to allow for identifying regularities beyond pure randomness in the study phenomena (e.g., by comparing experimental with control groups). This approach necessitates the statistical analysis of group-level distribution patterns. The statistical results, however, are commonly interpreted with regard to the single individuals (e.g., their beliefs). That is, from statistical analysis to result interpretation, psychologists shift their unit of analysis from the sample back again to the individual—without explanation but in line with their theoretical unit of analysis (Danziger, 1985; Richters, 2021; Uher, 2022b).

But in what ways can results on aggregates be informative about single individuals?

### 3.2.2 The ergodic fallacy: Psychology's common sample–to–individual inferences built on mathematical errors

Statistical analyses of aggregated data sets can reveal information about the single cases only when their synchronic and diachronic variations are equal (isomorphic)—a property of some stochastic and dynamic processes in non-living systems termed *ergodicity*. In the 1930s already, mathematical-statistical (ergodic) theorems[14] were used to prove that ergodicity does not hold for cases that vary, change and develop (Birkhoff, 1931). Hence, *psychology's study phenomena are non-ergodic*, which means that between-individual (synchronic) variations are uninformative about within-individual (diachronic) variations. Thus, when using sample-level analyses (e.g., factor analysis) to study individual-level phenomena (e.g., psychical 'mechanisms'), psychologists commit an inferential error—the *ergodic fallacy* (Bergman and Trost, 2006; Danziger, 1990; Lamiell, 2018, 2019; Molenaar and Campbell, 2009; Richters, 2021; Smedslund, 2016a, 2021; Speelman and McGann, 2020; Uher, 2022b, 2015d; Valsiner, 2014b; van Geert, 2011; von Eye and Bogat, 2006).

In sum, the higher-order complexity of psychology's study phenomena poses considerable challenges for empirical research. The uncontrollability of influencing factors requires statistical analyses of large samples. But individuals' complexity renders sample-level results uninformative about the single individual. These and further problems complicate the development of genuine analogues of measurement.

## 3.3 Psychological 'measurement' theories: Failure to conceptualise a coherent system of interrelated modelling relations

As Section 2 showed, measurement requires a coherent system of four interrelated modelling relations—each within a real and a formal study system and back and forth between them (arrows 1 to 4 in Figure 2). The 'measurement' theories established in psychology, however, such as Representational Theory of Measurement (RTM) and psychometrics, focus on just some of these modelling relations, thereby ignoring the overall model that is necessary to relate them coherently to one another.

### 3.3.1 Representational Theory of Measurement: Simple observable relations represented in mathematical relations

Representational Theory of Measurement (RTM; Krantz et al., 1971; Luce et al., 1990; Suppes et al., 1989) formalises axiomatic conditions by which observable relational structures can be mapped onto symbolic relational structures. It provides mathematical theories for this mapping (*representation theorem*), including permissible operations for transforming the symbolic structures without breaking their mapping relations onto the observable structures (*uniqueness theorem*; Narens, 2002; Vessonen, 2017). That is, representational theory specifies the semiotic representation of observable indications—the encoding and decoding relations in Rosen's structural model. The theory's focus on isomorphisms—thus, on reversible one–to–one relations between observables and data (arrows 2 and 4 in Figure 2)—presupposes that the objects of research feature properties with quantitative relations that are directly observable (e.g., 'greater than' or 'less than'). Such relations can be mapped straightforwardly onto a symbolic system that preserves these relations (e.g., ordinal variables; Suppes and Zinnes, 1963).

Psychologists, however, encounter tremendous challenges when trying to identify empirical regularities in observable (presumed) indications of psychical phenomena as well as (possibly) quantitative relations in these indications (e.g., in

14 The theorems were first derived in ergodic theory, a branch of mathematics originating in statistical physics.

behaviours, performances). Highly variable dynamic study phenomena necessitate fuzzy encoding relations, which can be defined and established differently. Specifying such many–to–one encoding relations is seldom straightforward. It requires theory-driven (in parts also arbitrary) decisions of what to formally represent and how. These decisions may impact the information encoded in the data—and thus, the results that can be obtained from them (Uher, 2019). All this further complicates the problem of measurement coordination in psychology (Luchetti, 2024; Uher, 2022b, 2023a). In all sciences, measurement requires highly selective and reductive representation. In psychology, it requires mapping information about a highly complex study system, which cannot be fully defined in principle (e.g., behavioural, psychical and belief systems), to a simple system, which can be fully defined (e.g., structural equation model).

Representational theory, however, provides neither concepts nor procedures for how and why some observations should be mapped to a symbolic relational system (Mari et al., 2017; Schwager, 1991). Concretely, it provides no concepts to specify the relations between observables and the (non-observable) quantity to be measured (measurand) in a study object. Nor does it provide concepts to specify the measurand's empirical interactions with the measuring instrument that first produce these observable indications (arrow 1 in Figure 2). Such specifications, however, are necessary to design suitable instruments and to operate them in defined empirical procedures (methods). They are also necessary to justify why some indications, but not others, should be observed—thus, to generate data that can be informative about the measurands (arrow 2). In view of this, it is unsurprising that representational theory provides no concepts for controlling the effects of influence properties and for modelling precision and uncertainty either. The theory confines empirical research to just simple observables that can be mapped easily onto useful mathematical relations, and vice versa. As Rosen (1985) highlighted, however, encoding and decoding (arrows 2 and 4) relations involve transformations that cannot be derived from within either system and that are therefore independent of these systems.

In sum, representational theory ignores the entire system of traceable modelling relations that must be coordinated and calibrated with one another to enable epistemically justified and publicly interpretable inferences from defined observable indications to the (non-observable) quantity of interest—the key criteria of measurement (Figure 2). Instead, it stipulates a purely representationalist and operationalist procedure that simplifies observations such as to align them to mathematically useful relations—in line with Stevens (1946, p. 667) earlier redefinition of 'measurement' as "the assignment of numerals to objects according to a rule" (other than randomness; Stevens, 1957). These simplistic notions formed the basis for psychology's theories and practices of pragmatic quantification and separated them from those of measurement used in metrology and physics (Mari et al., 2021; McGrane, 2015; Uher, 2021c). Still today, these representationalist and operationalist notions of 'measurement' underlie the psychology's main method of quantitative data generation—rating 'scales', in which numerical scores are straightforwardly assigned to specific answer categories.

These representationalist and operationalist notions of 'measurement' also underlie psychometrics—meant to mean the "science of measuring the mind" (Borsboom, 2005).

### 3.3.2 Psychometrics: Formal modelling aligned to statistical criteria and theories, enabling pragmatic result-dependent data generation

The triviality of the isomorphic relations in encoding and decoding (arrows 2 and 4 in Figure 3)—stipulated by representational theory and Steven's redefinition of 'measurement' and implemented in rating methods—shifted psychologists' focus away from the real study system (arrow 1) to the formal model (arrow 3). Statistical theories and methods, such as those of psychometrics, were advanced to develop sophisticated models and analyses that enable the reliable and purposeful discrimination between cases (e.g., individuals). This involved designing psychometric 'instruments' that allow for generating data with useful statistical properties (e.g., normal distribution, high item discrimination). Stevens' (1946) mathematically defined 'scales' (e.g., ordinal, interval, ratio)—although these are neither exhaustive nor universally accepted (Thomas, 2019; Uher, 2022a; Velleman and Wilkinson, 1993)—contributed further concepts to this end.

Psychometrics serves its pragmatic and utilitarian purposes well. But its approaches align the formal system (arrow 2 in Figure 3) to the criteria and theories on which the formal model and its manipulations are built (e.g., item-response theory)—regardless of the specific phenomena studied (e.g., behaviours and beliefs). Indeed, some even consider representation to be irrelevant for psychological 'measurement' (e.g., Borsboom and Mellenbergh, 2004; Michell, 1999). The epistemic necessity to conceptualise and implement an empirical interaction with the (non-observable) quantity to be measured in individuals gets out of sight. Psychometricians also overlook that identifying observable indications of these empirical interactions that may be informative about the measurands requires theoretical knowledge about *both* the real system studied and the methodical system (including the 'measuring instruments') used to study it (arrow 1). Instead, psychometricians choose 'instrument' indications (e.g., answer categories on rating 'scales') onto which pragmatically useful data structures (e.g., fixed numerical value ranges) can be mapped straightforwardly (Uher, 2018a, 2022a,b).

Hence, by focusing on statistical modelling (arrow 3, Figure 3), psychometricians neglect the three other modelling relations (arrows 1, 2 and 4) without which a formal system cannot be coordinated and calibrated with the real study system. Their interrelations are neither conceptualised nor empirically established but simply decreed, such as in the operationalist definition of 'intelligence' as what an IQ-test measures (Boring, 1923; van der Maas et al., 2014). Specifically, psychometricians fail to conceptualise the real study system—comprising the study object, the measurand, the instrument and their empirical interaction producing observable indications. Therefore, they overlook that the quantitative scores recorded in 'intelligence test' (e.g., number of correct answers) are properties of the *outcomes* of intellectual abilities but not of these abilities themselves.

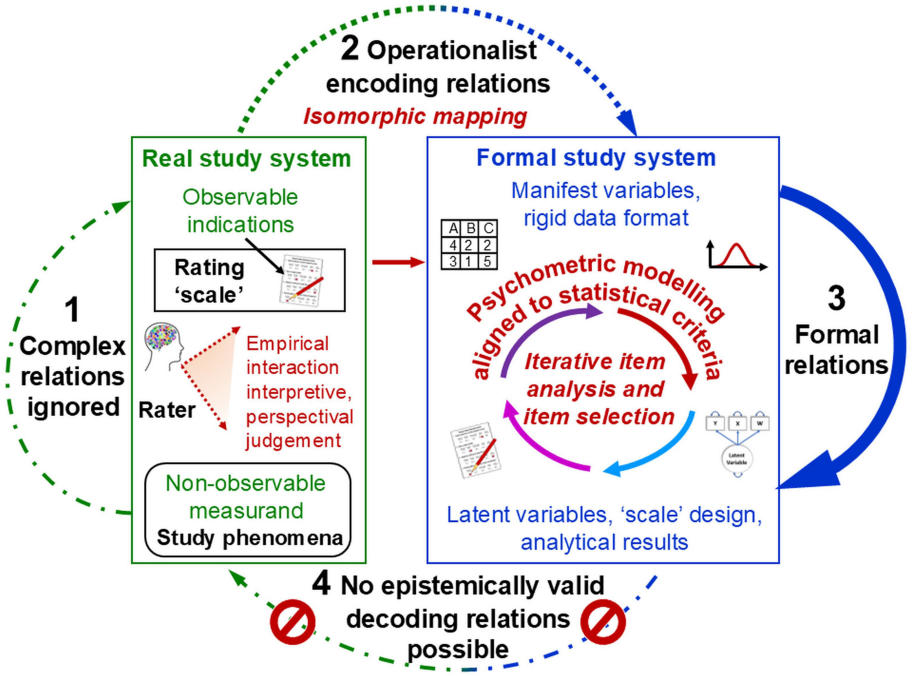



FIGURE 3
Psychometrics: Result-dependent methods of data generation and data analysis. Adapted from Uher (2022b, Figure 9).

Indeed, any test performance may involve several, qualitatively different intellectual abilities and modes of processing (e.g., symbolic, situational and verbal). More intelligent individuals may use *qualitatively* different (e.g., more efficient) abilities than less intelligent ones, different modes of processing and even multiple ones dynamically, leading to *quantitatively* different test performances. But none of these intricate many–to–one, one–to–many and many–to–many relations are considered in psychometrics. It only models relations of specific test outcomes to the abstract 'intelligence' construct that they operationally define, which is then re-interpreted as a real unitary object to be 'measured' (Khatin-Zadeh et al., 2025; Toomela, 2008; Uher, 2020b, 2021d,c, 2022b).

Psychometrics also provides neither concepts nor procedures for establishing unbroken traceable connections between results, measurands and instruments. As Section 2 showed, these are necessary to address the problems of circularity and coordination—thus, to provide evidence that a specific measurement procedure does indeed allow us to measure a specific property. Still, psychometric validity is often defined as "a property of measurement instruments that codes whether these instruments are sensitive to variation in a target attribute." This is "broadly consistent with the view that a test is valid if it measures what it should measure" (Borsboom et al., 2009, p. 135). Such causal measurand–result relations, however, are neither conceptualised nor empirically implemented. Therefore, the validity of psychometric 'instruments' can be analysed only regarding the coherence of their results with those obtained with other psychometric 'instruments' that are targeted at study phenomena that are considered to be theoretically related (Cronbach and Meehl, 1955).

These inconsistencies reflect the confusion between two incompatible epistemological frameworks, which is intrinsic to psychometrics. Psychometricians' declared aims and result interpretations invoke the *realist framework* of measurement. But psychometric theories and the implemented empirical practices are built on a *pragmatist utilitarian framework*. These pragmatic fundamentals are reflected, however, in validity concepts that focus on the results' practical use, such as their social and ethical consequences (Messick, 1995), or the inferences and actions that can be derived from them, such as regarding their plausibility and appropriateness (Kane, 2013; Uher, 2021d,c).

In sum, psychometricians focus on the formal model and its analyses (arrow 3) but neglect conceptualising and empirically implementing its interrelations with the real study system (arrows 1, 2 and 4). This aligns psychometric methods (e.g., 'scales') to statistical theories rather than to the study phenomena, thus enabling not traceable but *result-dependent data generation* (Figure 3) and leading to methodomorphism (Uher, 2020b, 2021c,d, 2022b, 2023a). This focus on statistical modelling abstracts away from the processes of measurement and thus, the actual study phenomena. It also obscures the data's meaning.

## 3.4 Psychology's focus on statistical modelling obscures the data's meaning

Psychologists' focus on statistics obscures the two distinct meanings that must always be conceptualised for empirical data—and thus, what these data actually represent. This highlights

peculiarities of sign systems that are crucial for enabling empirical science.

### 3.4.1 Statistics and algorithms: Analysing the syntax of data irrespective of their meaning with regard to the study phenomena—their empirical semantics

Statistics and other algorithms (e.g., data mining, machine learning) are formal methods that enable manipulations of formal models (arrow 3 in Figure 1), such as to identify regularities, interdependences, compatibilities or network structures in data sets. Statistical and algorithmic methods allow us to study how the data (e.g., variables, values) in a formal model are related to one another—their *syntax*. In linguistics, syntax denotes the set of language rules (e.g., grammar) that specify the structure and ordering in which words and phrases can be combined linearly to form sentences, which may influence their function in a sentence. Syntax allows us to indicate, for example, who is the actor of an activity and who the recipient. The words' meaning, in turn, arises from what they stand for and represent—their *semantics*. In linguistics, semantics denotes the set of rules that specify the meaning that words, phrases and sentences conventionally convey with regard to what they refer to (their referents). Thus, semantic meaning is established by way of a formal relation (Michaelis, 2003; Pattee, 2001).

The distinction between syntax and semantics is universal and basic to all life. In biophysics and biosemiotics, the DNA's *syntax* denotes the physical linear sequence of base pairs (copied into RNA through *transcription*[15]), whereas its *semantics* denotes the meaning that specific triplets on that sequence (codons) have for cells to instruct the production of specific amino acids (*translation*). That is, base triplets (codons) serve as physical tokens and carriers ("sign vehicles") that stand for something else (amino acids). What specifically they stand for is determined not physically (not by their molecular structure) but formally—on the basis of rules (described in the codon table; Abel, 2009, 2012; Pattee, 2021).

This illustrates the three distinct parts from which a *sign* emerges a whole (Figure 4). The *signifier* (e.g., a written word, an RNA codon) is the physical carrier that stands for something other than itself, which it represents, signifies or refers to—its *referent* (e.g., object, amino acid). The signifier's formal relation to a specific referent defines its semantic *meaning* (Ogden and Richards, 1923; Peirce, 1958; Rød, 2004; Uher, 2021c, 2022b; Vygotsky, 1962).

Hence, for empirical studies, we must always as*sign* both a syntactic and a semantic meaning to the signifiers that we use as data (e.g., variable names, numerical values). The *syntactic meaning* defines the data's relations within the formal system (Figure 5). Nominal, ordinal and ratio variable meanings, for example, define different mathematical relations for the same numerical values '1', '2' and '3'. Thus, *syntactically*, these data may denote categorical (qualitative) differences, order relations or quantitative relations in a model. The empirically established *semantic meaning*, in turn, anchors the data in that selected part of 'reality' that they are meant to represent and for which they serve as a surrogate to enable formal analyses (arrow 2). Thus, *semantically*, the same numerical data '1', '2' and '3' may refer, in different variables, to individuals' genders, shoe sizes, finger rings or hand gestures.

Statistics and other algorithms operate solely on the basis of a model's syntactic relations (arrow 3). They can neither establish nor analyse a model's relations to a real study system (e.g., genders, shoes or gestures). These methods perform purely syntactic data analyses *no matter what these data stand for* in a study—thus, regardless of their semantic relations to the real study system (arrows 2 and 4 in Figure 5). Ignoring the data's *empirical semantics* can lead to confusion about the syntactic relations that should be assigned to them (arrow 3) to appropriately match the empirical syntax of the real system (arrow 1).

### 3.4.2 Ignoring the data's empirically established semantics can lead to inappropriate syntactic (statistical) analyses

The data's semantic meaning is empirically established through encoding, which requires decisions about how to select and convert observations of elements of the real system into elements of the formal system (arrow 2 in Figure 5). To enable formal analyses, these conversion decisions must also consider syntactic relations that are identifiable in the selected indications of the real study system (arrow 1) and relevant to the research question (Uher, 2019). Qualitative differences (e.g., gender), rank-order differences (e.g., shoe sizes) or countable quantitative differences (e.g., finger rings) may be straightforwardly encoded into nominal, ordinal and ratio variables using isomorphic mapping relations as stipulated in representational theory. Mostly, however, psychologists encounter highly variable dynamic observables, such as in verbal and non-verbal behaviours (e.g., speech, gestures), that may be best described

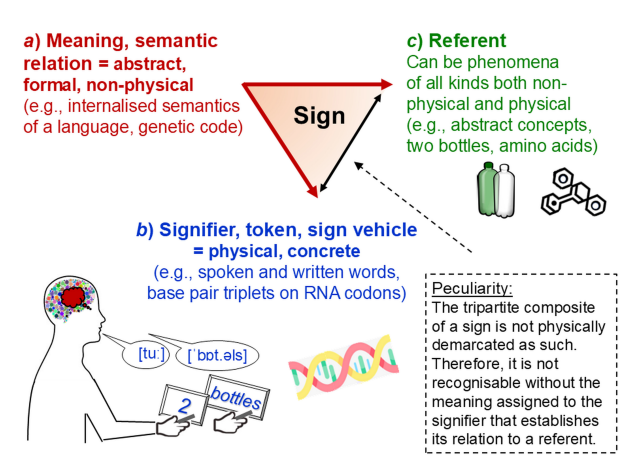



FIGURE 4
Signs: The meaning assigned to a signifier establishes its semantic relation to a referent. Adapted from Uher (2018a, Figure 3).

15 *Transcription* is the process whereby DNA is copied into RNA, following lawful (inevitable, necessary) pairings between bases (e.g., cytosine with guanine) determined by their molecular structure. *Translation*, by contrast, is the process whereby specific RNA codons instruct the synthesis of specific proteins following rules, which are not inevitable and not necessary given the bases' molecular structure but arbitrary—they could also be otherwise.

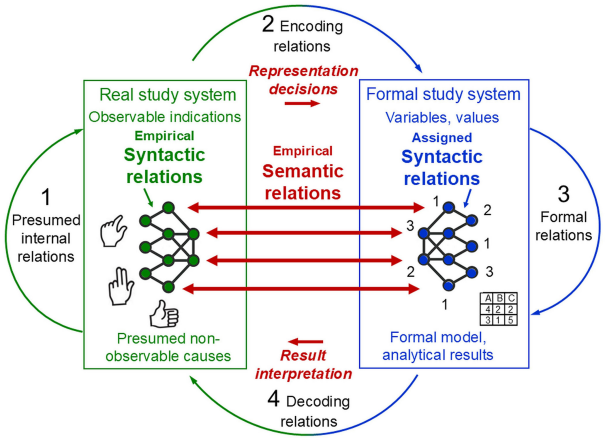



FIGURE 5
Data in formal models: Semantic and syntactic meanings.

in sets of fuzzy observables in which syntactic structures cannot be straightforwardly identified.

Necessarily, the syntactic relations assigned to the formal model (arrow 3 in Figure 5) are also informed by the formal manipulations that they enable (e.g., statistical analysis). But because the model is just a surrogate, its syntactic relations must be aligned to those that are identifiable in the observables of the real study system (arrow 1). This is crucial because observations constitute the only direct empirical evidence that can be obtained about the real study system. Observational raw data form the basis for modelling, in the formal system, the (presumed) non-observable relations in the real study system (for which different syntactic relations may be conceived) as well as for testing the model's appropriateness through coordination and calibration. Importantly, which observable indications and which of their syntactic relations are (possibly) informative about the non-observable measurands depends not on the indications' ease of observability but on the theories about the objects of research, the measurands, instruments and their empirical interactions. Selecting indications by desirable syntactic structures, as done in psychometric 'instrument' design and stipulated by representational theory, leads to methodomorphism, result-dependent data generation—and eventually to biased 'measurement' results.

Hence, whether a model's syntax and the statistical analyses performed on it (arrow 3 in Figure 5) are appropriate for, and thus informative about, the *empirical syntactic relations* in the real study system (arrow 1) depends on the model's empirically established relations to that real system (arrows 2 and 4). Ignoring the model's *empirical semantic relations*, such as by neglecting encoding, coordination and calibration, can lead to logical errors and inappropriate data analysis. For example, students sometimes analyse means and standard deviations for data on persons' gender (encoded, e.g., as '1', '0' or '1', '2', '3') by taking them for ratio rather than nominal values. Thus, they ignore their empirical semantics of their data, established during data generation, and assign a different syntax to them. Syntactic mismatches between real and formal system also occur when researchers encode the verbal answer values of Likert 'scales' ('instrument' indications) in numerical scores and assign to them desired syntactic relations (e.g., order, interval). This ignores the empirical semantics that the researchers themselves establish by making these assignments. Specifically, what justifies the assumption that "agree" (encoded as '4') reflects more than "disagree" (encoded as '2')? How can we assume that "neither disagree, nor agree" (encoded as '3')—thus, having no opinion or finding the item not applicable—constitutes more than "strongly disagree" (encoded as '1')? Given the verbal answer categories' logico-semantic meanings, it is no wonder that raters interpret these not as reflecting order or interval relations but only as categorically—thus, qualitatively (nominally)—different (Uher, 2018a, 2022a, 2023a).

In sum, *data always have, at once, semantic and syntactic relations*. Their semantic relations are established through coordinated empirical relations to the study phenomena. These determine which syntactic relations can be assigned to the data to appropriately represent those identifiable in the real system, thus also enabling their calibration.

Establishing the data's empirical semantics is complexified by human language. Its peculiarities first enable the use of language-based methods in empirical research, but they also obscure psychology's measurement problem.

## 3.5 Natural language: Intuitive and ease of use obscures inherent complexities and common confusions

Language is an essential means for psychological research because psychical phenomena (e.g., thoughts and beliefs) are accessible only by the individual itself, and they can be accessed in others (e.g., research participants) only through language (Uher, 2016a; Valsiner, 2007; Vygotsky, 1962). In everyday life, we use language to exchange with others intuitively and without much reflection. Yet, this ease of use often leads us to overlook unparalleled complexities that challenge empirical research, especially measurement.

### 3.5.1 Language and mind: Different yet inseparable systems

Language, as we have seen, is a complex sign system. It involves physical carriers (signifiers; e.g., spoken or written words) that stand for and refer to something else (referents), which establishes their semantic meanings. The rules underlying the semantics and syntax (e.g., grammar) of language are construals of human minds. This also applies to pragmatics, the rules specifying the language's function in the context of social interaction (e.g., the communicating persons' intentions and beliefs). These rules enable competent language users to express complex meanings with some flexibility and in context-dependent ways as well as to infer the specific meaning that others may want to express verbally in a given context. The language rules established in a community feed back to the individuals who develop and use them by mediating and shaping both their intra-individual and inter-individual processes (e.g., feeling, thinking, memorising, interacting and negotiating) as well as the social institutions aimed at regulating these processes (e.g., family and government). Therefore, language and psyche are inseparable from one another while still constituting different kinds

of phenomena (Peirce, 1958; Uher, 2015b,a, 2016a, 2018a; Valsiner, 2000, 2014a; Vygotsky, 1962).

We use our maternal language effortlessly and without being fully aware of its *inbuilt semantics, syntax and pragmatics*. This is because these complex rules form an inherent part of our psychical systems after we internalised them as children during our language socialisation. Therefore, as native speakers, we often struggle to explicate the rules that we intuitively use, and we are often surprised what rules foreign learners of our language can state. That is, we are *competent without comprehension* (Arnulf, 2020; Dennett, 2012). This entails that we rarely become aware of the inherently representational nature of language, which is built into its semantics. Indeed, in our minds, we do not perceive our words just as tokens of the objects to which they refer but as these objects themselves. This illusion makes language so highly functional in everyday life. Yet, it becomes apparent again in our struggles of learning a foreign language when we have to acquire new words as arbitrary tokens to refer to the things of the 'world'. But once we have internalised (at last parts of) a given language's inbuilt semantics, we cannot easily blank it out anymore to enable reflection and reflexivity about the ways in which it modulates and shapes our thinking. This is what makes naming a word's font colour more difficult when that word itself denotes another colour (Stroop effect)—unless we do not know the language, then the task is easy.

Therefore, we often forget that semantic relations are just in our minds, linking our words and thoughts seamlessly with the objects to which they refer. As Alan Watts stated:

> "When I use the word *thinking*, I mean precisely that process of translating what is going on in nature into … symbols … [U]sing symbols and using conscious intelligence has proved very useful to us. It has given us such technology as we have; but at the same time, it has proved too much of a good thing. At the same time, we've become so fascinated with it that we confuse the world as it is with the world as it is thought about, talked about, and figured about—the world as it is described. The difference between these two is vast…" (italics as in original; Watts and Watts, 1996, p. 26).

Our ability to use the inbuilt semantics of our natural language intuitively and with ease, lets us often overlook its representational nature and confuse our words with 'reality'.

### 3.5.2 The map is not the territory, the model is not reality, the word is not the thing

Korzybski (1933) established *general semantics*—the study of language as a representation of 'reality'. In his critique of traditional assumptions about language, he illustrated the distinction between a real object and its formal representation by stating that

> "A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness" (Korzybski, 1933, p. 58).

Korzybski used the map–territory relation to illustrate the distinction between our perceptions or beliefs of something and the actual 'reality' of it. Specifically, a map of a city is not that city in itself. Reading a map is not the same as walking the streets. Maps depict in abstract symbolic ways only those parts of a territory that are seen as relevant for some purpose (encoding, arrow 2 in Figure 1). Therefore, we can establish for the same territory different maps (e.g., road maps, geographical or political maps). That is, maps are reduced semiotic representations. All maps are limited. They may be incomplete or outdated. 'Reality' may change (e.g., closed roads). Moreover, using maps requires interpretation (decoding, arrow 4), which may involve errors. Therefore, our maps of some 'reality' (arrow 3) neither are that 'reality' in itself (arrow 1) nor can they exactly match that 'reality' (arrows 2 and 4).

Korzybski (1933) highlighted that we tend to mistake our conceptual models of 'reality' for that 'reality' in itself. This occurs when we ignore that the *word is not the thing*, the *abstraction of something is not that something in itself*—and thus, also, that the *theory is not what it describes* and that the *data are not the study phenomena* for which they stand (Uher, 2018a, 2021a, 2022b). As Alan Watts put it more vividly,

> "symbols bear the same relation to the real world that money bears to wealth. You cannot quench anybody's thirst with the word 'water' just as you cannot eat a dollar bill and derive nutrition." "Money simply represents wealth in rather the same way that the menu represents the dinner[16]" (Watts and Watts, 1996).

Korzybski warned of the logical fallacies that ensue when the model is mistaken for 'reality'. These occur not just in everyday life but also in science. In psychology, for example, latent variables that were statistically derived in a formal (e.g., factor analytical) model (arrow 3 in Figure 3) are often interpreted as 'traits', 'psycho-physical mechanisms' or 'personality factors' that causally underlie individuals' behaviours, thoughts and feelings (arrow 1; Uher, 2013, 2018b, 2022b). In psychological jargon, the term 'data' is often used to denote both the study phenomena (e.g., in individuals) and the formally encoded information about them (e.g., on spreadsheet; Uher, 2021a). The term 'variables', as well, often denotes not just parts of formal models but also the modelled real objects themselves (Danziger and Dzinas, 1997; Maraun and Gabriel, 2013; Maraun and Halpin, 2008; Uher, 2021d,c). The confusion of the model with 'reality' is also reflected in the notions that we would study 'correlated behaviours' or 'measure variables' as well as in the demand to grant "a serious ontological status to variables" (Borsboom, 2008, p. 41). Conflated jargon promotes such confusions because it leads researchers to neglect a formal system's empirically established semantics, which defines its relations to the real system—and thus, these systems' epistemic separation.

In sum, language and its conventional rules are construals of human minds, which, at the same time, mediate and shape individuals' psychical processes. Its intuitive and ease of use enables but also obscures its inherently representational function, leading to common confusions between words and the 'reality' that they denote. When using language-based 'instruments', these challenges are incorporated directly into psychological 'measurement'.

16 The menu–food metaphor was also used by Arnulf et al. (2024).

## 3.6 Language-based 'scales' obscure the measurement process

Psychological 'measurement' is unthinkable without everyday language. It relies on the idea that any phenomenon of interest can be empirically studied, and even 'measured', as long as it can be verbally described. Accordingly, rating 'scales' comprise brief verbal descriptions of the phenomena with which *raters*—the persons using these 'instruments'—are assumed to interact. While efficient and easy to use, modelling this process is intricate.

### 3.6.1 Obscured distinctions between psychical phenomena, language-based 'instruments' and formal models

In physical measurement, all elements of the real study system—the objects studied, those used as measuring instruments, their lawful empirical interactions and the indications thus-produced—are all of physical nature. The model that semiotically represents selected information about them, however, is formal, thus non-physical (Figure 2). In psychology, by contrast, real and formal system cannot be easily distinguished. Psychical phenomena (e.g., intellectual abilities, beliefs) are non-physical, abstract and represent information—just as the formal models developed about them. Language, here used as method and 'instrument', is a complex sign system to communicate information—thus, a formal system as well. These peculiarities complicate the epistemically necessary distinction between the real and the formal study system. It also blurs, within the real system, the distinction between the phenomena studied and those used as 'instruments' for studying them. This complicates the conceptualisation of how the 'instruments' can be used in a given method to produce information about the study phenomena—*psychology's measurement problem*.

These epistemically necessary distinctions are further hindered by the ambiguous use of the term 'scale' in psychology. On the one hand, it refers to Stevens' (1946) concept of 'measurement scales' which defines variables with specific mathematical properties (e.g., ordinal, interval and ratio)—thus, structures of formal models. On the other hand, the term 'scales' denotes the 'instruments' that enable empirical interactions with the measurands, just like physical measuring devices (e.g., weighing scale; Uher, 2022a). Formal scale and physical scale, however, although coordinated and calibrated with one another, are epistemically distinct elements of measurement (Figure 2). In psychology, this distinction is obscured when the rating items serve both—as descriptions of the study phenomena in the verbal 'scale' and as item variables in the formal model (Uher, 2018a). But raters interact only with the item statements of the 'scales', not with the statistical models through which these were designed. So, what function do the item statements have when used as 'instruments'?

### 3.6.2 Specifying the phenomena to be 'measured' through the inbuilt semantics of everyday language: Collective fields of meanings

Rating items categorise and describe the phenomena to be 'measured'. Worded in everyday language, this enables lay persons to use rating 'scales' with just minimal instruction and without any training. This differs fundamentally from many kinds of physical and behavioural measurement (Uher, 2018a, 2021c). Thus, psychologists capitalise on raters' and their own intuitive knowledge and use of natural language and its *inbuilt semantics*.

The inbuilt semantics of our words—their *conventional meanings*—are described in our dictionaries. Words can be grouped by their dictionary meanings and described in their semantic relations with other words using logic-based formalisms. These interrelations between words form *semantic networks*, which can be visualised in graphical networks. These networks describe common structures in the organisation of knowledge representations and information retrieval pathways that are socially shared by competent users of a given language (Arnulf et al., 2018; Pirnay-Dummer et al., 2012). In the semantic space of a language, a word's multi-dimensional associations with other words span a *field of meaning* (Rosenbaum and Valsiner, 2011; Uher, 2018a, 2022b, 2023a).

The general semantic meaning that language users *collectively* construe for a word is derived and abstracted from the specific meanings that individual users *locally* construe for it in the specific contexts of its use. A 'house', for example, may mean a building serving as family quarters, refuge or shelter, but also a dynasty (House of Windsor), governmental institution (House of Commons), gathering place for specified activities (coffee house), or a business organisation (publishing house). That is, words may refer to concrete observables (e.g., buildings)—thus, they have a primary literal meaning (*denotation*). But many words also often imply *interpretations and explanations* of their referents (e.g., regarding their purpose) or they may be used as metaphors (e.g., 'house' as 'dynasty'; Lakoff and Johnsen, 2003). Thus, words may also have additional non-literal meanings (*connotations*). These meanings are more abstract and socio-culturally construed and often cannot be easily traced back anymore to their formerly concrete references and contexts (Deutscher, 2006).

This also applies to psychology's study phenomena. Most behaviours possess various observable features and can therefore be interpreted differently regarding possibly associated psychical phenomena (e.g., different intentions or feelings; Shweder, 1977; Smedslund, 2004; Toomela, 2008; Uher, 2015d). Describing the act of taking an object as 'finding', 'exploring', 'securing', 'catching', 'seizing', 'grabbing' or 'stealing' implies different interpretations regarding the actor's (presumed) goals and intentions in the given context. That is, behaviours can be described in their momentary and localised physical properties (Uher, 2016b). But their explanations can go well beyond the here-and-now and can invoke various interpretive perspectives. These all follow logical principles (Kelly, 1955; Smedslund, 2004) yet without being logically determined by the behavioural act itself (Shweder, 1977).

Many words also imply *normative evaluations*. As members of the same community, individuals are substantially similar to one another. Evaluating normativity therefore requires abstracting from commonalities and focussing instead on minor variations (e.g., behavioural, facial) that are informative for differentiating between (groups of) individuals. Promoted by social appraisal (e.g., valued, sanctioned) and putative explanations (e.g., innate, intentional), socially relevant variations are often exaggerated. Then they appear in people's minds to be larger than they

actually are, thereby acquiring *salience* (Uher, 2013). Those salient variations that are considered most important in a language community may eventually become encoded in words (lexical hypothesis; Allport and Odbert, 1936; Galton, 1884).

All this entails that everyday language is replete with inferential assumptions, implicit connotations, socio-cultural valences, interpretations and putative explanations (Shweder, 1977; Smedslund, 2004). This allows rating items to be worded such that they refer to a broader range of phenomena and contexts that raters could consider as well as to capture raters' interpretation, explanation and normative appraisal of them (Uher, 2015c, 2018a, 2023a). This shows again key differences between physical measurement and the pragmatic quantifications used in psychology. They arise from the fact that psychology's focus is on the individual (subjective) and socio-cultural (inter-subjective) interpretations, explanations and appraisals of observable (and inferred non-observable) phenomena—thus, on the meanings that these have for individuals and communities. This differs from physics and metrology, which aim to explore just the phenomena and their relations in themselves but not also our human experience and apprehension of them (Uher, 2020b, 2021b; Wundt, 1896).

This also highlights the crucial role of persons in the use of rating 'scales'.

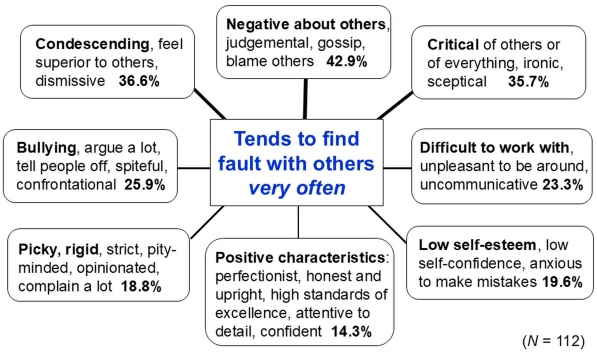



FIGURE 6
The inbuilt semantics of rating 'scales': collective field of an item's local context-specific meanings. The collective general meaning of the item "tends to find fault with others", used to operationalise the construct 'Agreeableness' in the BFI-10. Its field of meaning is illustrated through the main themes that summarise the local context-specific item meanings that $N = 112$ participants construed for this item, described in their own words in terms of behaviours that a fictitious person scoring high on the item (indicated by 'very often') would typically show. Percentages indicate the proportions of participants providing interpretations that are pertinent to a given theme (multiple nominations per person possible).

### 3.6.3 Psychology's measurement problem is left to raters' intuitive decisions and local interpretations of standardised rating 'scales'

Physical measurement requires objects used as measuring instrument that lawfully interact with the objects of research, thereby producing an indication from which information about the object's measurand can be derived. By contrast, language-based 'instruments' themselves cannot interact with anything. Language involves not lawful relations but rules, which must be known and applied by persons. That is, *language-based methods require interpretation*, which is always context-specific, and thus variable. Moreover, psychology's objects of research are (primarily) human beings and specific phenomena and properties that are accessible only by persons (e.g., intensity of feelings, strength of beliefs) or that are studied from their individual perspective (e.g., perceived frequency, ascribed intentionality or normativity of others' behaviours). Therefore, it requires persons (e.g., research participants, patients) to interpret and use rating 'instruments' and to identify relevant study phenomena. These persons must also interact with and judge these phenomena for specific purposes and from specific interpretive perspectives, and they must visibly indicate the outcomes thus-produced on the rating 'scale' (e.g., by ticking a box). Hence, in psychology, the real study system involves *complex interactions* that are executed by persons. These persons therefore play a crucial epistemic role in the 'measurement' process.

Language-based 'scales' are standardised through identical wordings of items and answer categories and are therefore often thought to mean the same for all raters. This implies the assumption that all individuals interact with these 'scales' in the same ways and produce indications that have the same meaning for everyone. But from the entire field of an item's general meaning, raters construe only a specific one that matches the context and specific interpretive perspective that they consider for a rating. Thus, they construe a *local meaning*. Figure 6 summarises the local meanings that 112 research participants independently construed for the item "tends to find fault with others" from a popular 'personality' 'scale'[17]. It depicts the broad field of meaning that this item statement collectively had for all raters but also the diversity of local meanings that they considered individually (Uher, 2018a, 2023a). That is, some raters read the item as "condescending," others as "being picky, rigid," still others as "having low self-esteem" or "being perfectionist, honest and upright" (Figure 6). On average, each rater considered only two different item meanings ($M = 2.08$; $SD = 0.92$; range = 1 to 5). No one considered the entire field of meaning. Thus, when used empirically by raters, standardised rating items have no unitary meanings.

Such variations in item interpretation, which occur both between and within individuals, were demonstrated also for other items of the same questionnaire (Uher, 2018a, 2023a) as well as in other studies (e.g., Arro, 2013; Lundmann and Villadsen, 2016; Rosenbaum and Valsiner, 2011; Uher and Visalberghi, 2016; Valsiner et al., 2005; Wagoner and Valsiner, 2005). The general dictionary meaning of rating items—their *inbuilt semantics*—can also be studied with artificial intelligence technologies.

### 3.6.4 The inbuilt semantics of rating 'scales': Natural language processing algorithms reveal its use by raters for mental short-cuts

*Natural language processing (NLP) algorithms* are types of artificial intelligence (AI) technologies to computationally analyse and process human language data. They are used either to identify specific structures and explicit rules in texts ('understanding')

17 Big Five 10-item short version (BFI-10; Rammstedt and John, 2007).

or to produce texts from the algorithms identified ('generative'). NLP algorithms dissect textual data sets (corpora) using statistical, mathematical or probabilistic methods (e.g., machine learning techniques). They analyse sentence structures (syntax) and keywords in order to identify or reproduce patterns and relations between words in sentences. NLP algorithms can be used, for example, to correct spelling (autocorrect), predict the next word given the preceding words (autocomplete), convert spoken words into written text (speech recognition), translate text from one language into another (machine translation) or extract the possible meaning (inbuilt semantics) of a sentence from its keywords and context or from the words' dictionary-based interpretation (content categorisation, automated text summarisation). To enable this, some NLP algorithms also rely on well-defined semantic and knowledge representations that are taken from linguistically established (previously hand-coded) dictionaries (Khurana et al., 2023). That is, NLP algorithms can formalise structures and explicit rules that underlie a given natural language and that can use these to analyse and generate texts.

Analyses of popular rating 'scales' with NLP algorithms showed that the overlap in their items' inbuilt semantic meanings explained 60%−86% of the variance commonly found in ratings empirically obtained on these items (e.g., using factor analysis; Arnulf and Larsen, 2015; Arnulf et al., 2014). This sheds a new light on psychometrically established nomological networks. Traditionally, these are interpreted as sets of correlating item variables that encode the observable indicators (e.g., specific behaviours) through which a construct is operationalised (e.g., a 'trait'). Instead, nomological networks may also largely reflect just the inbuilt semantic networks underlying the items' general (dictionary) meanings rather than any empirically derived structure in the phenomena described (Arnulf et al., 2024). Hence, ratings may reflect *likeness in semantic meaning* rather than *co-occurrence likelihood* of the phenomena described (Shweder and D'Andrade, 1980).

This was also demonstrated in multi-method studies. Associations of observer ratings on behaviour-descriptive items reflected their inbuilt semantic meanings but not the empirical patterns by which the described behaviours actually occurred in the same target individuals. Indeed, time-based measurements of functionally similar behaviours (e.g., different acts of aggression) showed only low to moderate internal consistency but substantial temporal consistency, thus indicating individual specificity ('personality'). Observer ratings of the same target individuals on items describing the same behaviours, by contrast, were internally consistent—in line with their inbuilt semantic meanings (Uher et al., 2013a; Uher and Visalberghi, 2016). Thus, the inherently interpretive perspectives of rating items, reflecting socio-culturally ascribed valences and normativity, may influence and even bias perceptions and judgements of their observable referents (Shweder, 1977; Uher, 2022b; Vygotsky, 1962).

All this suggests that raters may use the inbuilt semantics of rating items also as mental short-cuts to simplify their rating task. Specifically, as thinking and learning agents, many raters do not fail to notice that rating 'scales' commonly contain, in randomised order, items with similar content (a necessity for the psychometric analyses). Therefore, raters may focus on a few salient referents just for the first items on a 'scale'. For any items perceived as 'repetitive', however, they may generate their responses more efficiently by focussing just on their inbuilt semantic similarity instead of construing local meanings and considering specific referents for each single item anew (Uher, 2015c; Uher et al., 2013b).

Raters' locally construed meanings are commonly not inquired, however. Therefore, it remains unknown which specific phenomena and contexts they have considered in a given rating, from which specific perspectives (e.g., normative appraisal) they have judged them, and how they actually used the item 'scales'. In consequence, the distinction between the objects studied and the objects used as 'measuring instruments' *is left to intuitive decisions of raters,* who are commonly lay people. The intricate problem of conceptualising how the methods (and 'instruments') interact with the study phenomena and can provide epistemically justifiable information about these phenomena—*psychology's measurement problem*—therefore remains undefined and unexplored (Uher, 2022b, 2023a). This ultimately obscures also the relations of the data and the formal model to the real phenomena under study.

### 3.6.5 Researchers' focus on the inbuilt semantics of rating 'scales' obscures the data's empirical semantics and syntax

Given that only the raters know how they have interpreted and used a 'scale', only they can know what the rating data ultimately stand for and refer to—their *empirical semantics*. When encoding and analysing rating data, however, psychologists consider only the items' general meanings—their *inbuilt semantics*—ignoring the fact that raters consider for the *same* item *different* local meanings, *different* specific phenomena, *different* contexts and *different* interpretive perspectives. These *one–to–many relations in the data's empirical semantics* preclude tracing the data back to the real phenomena and contexts that raters have considered and judged and that their ticks on the 'scales' were meant to indicate. But because raters' decisions are commonly not inquired—despite their crucial role in the data generation—these *breaks in data generation traceability* remain undetected (Figure 7).

Moreover, raters cannot indicate the outcomes of their interactions with the study phenomena (their judgements) in ways that they deem suitable for communicating them. Instead, raters can indicate their judgements only in a bounded set of verbal response categories that are specified a-priori by the researchers. We already discussed the syntactic mismatches that occur in agreement (Likert) 'scales' between raters' primarily qualitative interpretation of 'scale' categories (given their inbuilt logico-semantic meanings) and researchers' numerical encoding of them. Syntactic mismatches can also occur in frequency 'scales' when raters are forced to use the *same* 'scale' for *different* items—*regardless of the phenomena described*. Because different phenomena generally occur at different rates (e.g., chatting vs. shouting), this requires raters to indicate a broad range of quantities *flexibly* in the same 'scale'. Raters can do so only by assigning *different* quantitative meanings to the *same* response value—a necessity that violates core ideas of measurement (Uher, 2022a). These *syntactic many–to–one relations* preclude that raters' indications on the 'scale' can be traced back to the syntactic

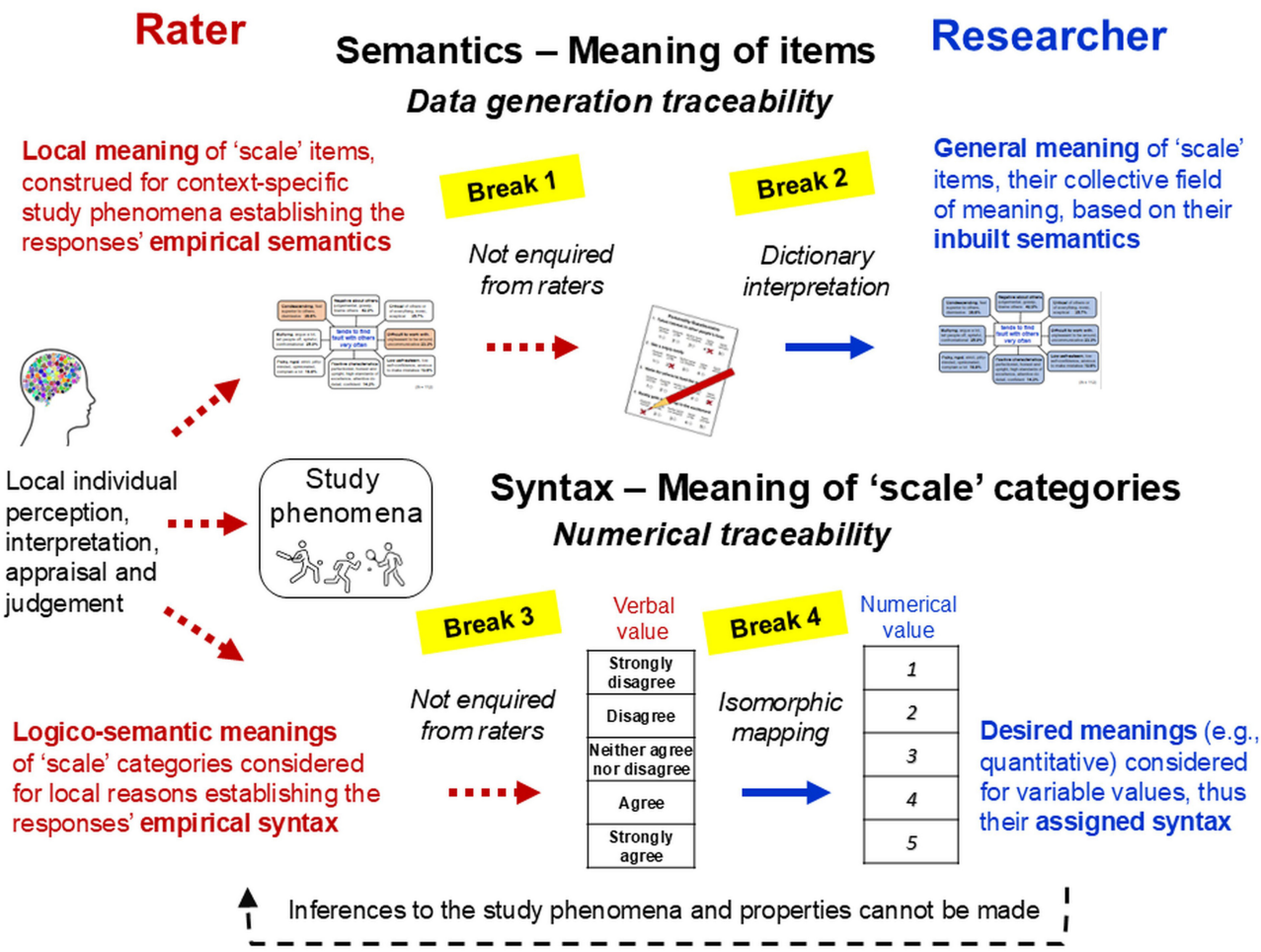



FIGURE 7
Four-fold break in data generation traceability and numerical traceability obscures the data's empirical semantics and syntax. Based on Uher (2018a, Figure 15).

relations that they actually considered in the study phenomena. But these *breaks in the numerical traceability* of rating data remain undetected when raters' rationales for ticking 'scale' boxes (indications) are not inquired and researchers consider instead only the syntactic relations that they themselves assign to the 'scale' categories and their numerical encodings in the data (Figure 7).

In sum, using language-based 'scales' to generate numerical data introduces several breaks in the semantic and syntactic relations between real and formal study system. But these breaks go unnoticed because quantitative psychologists do not consider raters' local interpretation and use of item 'scales' but rely instead solely on the items' *inbuilt semantics* and on the syntax that they, as researchers, assign to raters' numerically encoded responses. Intuitive reliance on the inbuilt semantics of language-based methods also obscures the epistemically necessary distinction between the actual study phenomena and their verbal descriptions on the 'instruments' and leaves it to raters' intuitive unknown decisions. In consequence, researchers cannot assess if their own decisions about how to encode raters' responses in numerical data (arrow 2; Figure 3) are appropriate (e.g., logical, consistent) for the real study phenomena. Researchers also cannot assess if their statistical analyses of the thus-generated data (arrow 3) as well as their interpretations of the results obtained are semantically and syntactically appropriate for the real study system (arrow 4) and can reveal epistemically justified information about its internal relations (arrow 1). That is, psychology's standard practice of generating quantitative data with rating 'scales' fails to empirically establish the system of interrelated modelling relations that is required for measurement.

## 4 Statistics and language-based methods in quantitative psychology: Implications and future directions

Language is human's greatest invention (Deutscher, 2006). With words, we can refer to objects of consideration even in their absence (meaning), and although what we say or write (signifiers) typically bears no inherent relations (e.g., resemblance) to the objects referred (referents). This representational function

of language—built into its semantics—is internalised in our minds and fundamental to our abstract thinking. However, we do not perceive our words just as tokens of the objects to which they refer but as these objects themselves. In our minds, we therefore easily mistake the word for the thing, the map for the territory, the menu for the food—the 'world' as it is with the 'world' as it is thought about and described. This also misguides our scientific thinking at times and leads to fundamental errors.

## 4.1 Psychologists' cardinal error: Failure to distinguish the ontic study phenomena from the epistemic means of their exploration

Our tendency to mistake verbal descriptions for the phenomena described affects psychology in particular ways because we can access others' psychical phenomena never directly but only mediated through language. Unawareness of its inherently representational nature—its *inbuilt semantics*—often obscures the epistemic necessity to distinguish the study phenomena (e.g., raters' thoughts or feelings) from their verbal description in the language-based methods used for exploring these phenomena (e.g., item 'scales', variable names). Failure to make this crucial distinction thus *confuses ontological with epistemological concepts*—therefore termed *psychologists' cardinal error* (Figure 8; Uher, 2022b, 2023a).

Psychologists' cardinal error can occur in various parts of the empirical research process. This logical error makes the distinction of disparate research elements and activities technically impossible and distorts basic concepts, methods and result interpretations (Uher, 2022b, 2023a)—such as in the processes required for measurement.

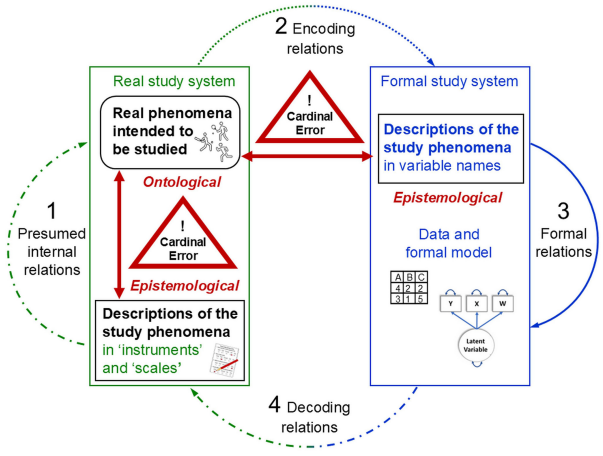



FIGURE 8
Psychologists' cardinal error: Failed distinction between ontological and epistemological concepts promoted by the inbuilt semantics of language-based methods.

### 4.1.1 The inbuilt semantics of language-based methods obscures the distinction between the 'instruments' and the phenomena to be studied

The failure to conceptualise measurement processes in many psychological studies is often due to psychologists' cardinal error. This is because, when considering only their items' *inbuilt semantics*, psychologists often fail to distinguish the study phenomena's descriptions that are used as 'instrument' from the described phenomena themselves that are intended to be studied (Figure 8). This error often underlies evaluations of face validity and content validity of psychometric 'instruments'. It also underlies the widespread belief that any rating 'scale' that is *nominally (by name)* associated with a study phenomenon could be a valid method for empirically studying it (e.g., 'neuroticism scale'). This *nominalism* and *toolbox thinking* contribute to the proliferation of overlapping rating 'scales' (e.g., various anxiety 'scales') and of the likewise overlapping constructs that their items are meant to operationally define (Sechrest et al., 1996; Toomela and Valsiner, 2010; Uher, 2021b, 2022b).

### 4.1.2 Mistaking judgements of verbal statements for measurements of the phenomena described: The risk of pseudo-empirical research

Psychologists' cardinal error also occurs when, through the *inbuilt* semantics of item 'scales', researchers intuitively establish—in their minds—relations from their 'instruments' to the study phenomena described. Their (and raters') mental construction of these relations (meanings) is necessary to specify the phenomena (referents) to be considered. But these mental relations only pre-structure their thinking—they *do not, themselves, implement any empirical relation to the real 'world'*. Yet, because these relations are thought, they obscure the necessity to scrutinise what empirical connections are actually implemented in a study—and thus, what *empirical semantics* are established for the data thus-produced. 'Personality' ratings, for example, enquire about habitual behaviours, which have necessarily already occurred in the past. Past events can be mentally (re-)construed. But traceable empirical interactions with those events, as required for measurement, can no longer be implemented.

In this way, the *inbuilt semantics* of language often leads researchers to misinterpret raters' judgements of verbal statements as measurements of the phenomena described in those statements (Figure 8). The necessity to conceptualise and empirically implement a coordinated and calibrated system of four interrelated modelling relations, as described in Rosen's general process scheme (Figures 1, 2), gets out of focus—and with it the actual phenomena under study. This entails the risk of replicating just verbal descriptions rather than exploring the real phenomena for which these are meant to stand (Baumeister et al., 2007; Cialdini, 2009; Doliński, 2018; Osborne-Crowley, 2020; Teigen, 2018; Uher, 2022b, 2023a; Wojciszke and Bocian, 2018). This puts quantitative psychology at risk of doing *pseudo-empirical* research, which mostly re-discovers what is necessarily true given the logico-semantic relations built into its language-based methods (Arnulf et al., 2024; Shweder, 1977; Shweder and D'Andrade, 1980; Smedslund et al., 2022; Smedslund, 1991, 2016b).

#### 4.1.3 Advancing just statistical methods and models: Creating a formal sphere disconnected from the 'reality' to be explored

The focus on statistics leads quantitative psychologists to create formal spheres in which ever more sophisticated analyses and models can be developed. In the formal 'world', there are no limits. This, however, ignores the epistemic necessity to empirically connect the formal models and data with the real study system, for which they serve only as surrogates—thus, to establish their *empirical semantics*. But the *inbuilt semantics* of the language terms that are used as data and variables in statistical models often lead psychologists to mistake the data for the phenomena and the models for the 'reality' described—thus, to commit the cardinal error of confusing epistemological with ontological concepts. This confusion creates a data 'world', a parallel universe of purely verbal representations but that has no traceable connections to the real 'world'. Quantitative psychology then becomes a mere data science.

This empirical break leads many psychologists to overlook that *low replicability* is not just an issue of *epistemic uncertainty*, which could be remedied with more sophisticated procedures, but that it also reflects the study phenomena's *ontic indetermination*, variability, changeability and developmental nature. Psychology must advance concepts and empirical practices that are adapted to and appropriate for these peculiarities rather than focus only on what is possible in purely formal (e.g., statistical) systems. We cannot indulge in ever more complicated formal manipulations that have no counterparts in the 'reality' that we aim to explore because this entails a proliferation of theories, constructs and supposed psychical phenomena for which there is little or no actual evidence. Ever more complicated statistics and their meticulous and transparent application (e.g., open science) therefore cannot tackle psychology's crises (e.g., in replicability, validity, generalisability), as currently believed, and but will only exacerbate them (Kellen et al., 2021; Uher, 2021b, 2022b; Uher et al., 2025).

#### 4.1.4 Statistics is not measurement: Psychology's pragmatic quantifications are numerical data with predictive power but without explanation

The common belief that statistics constitutes measurement is not just unwarranted. It is also misleading. In both everyday life and science, the term measurement implies that some part of 'reality' is being quantified (e.g., some apples' weight). Measurement results are regarded as epistemically justified (e.g., we trust the shops' calibrated weighing scales; criterion 1) and publicly interpretable regarding their specific quantitative meaning for the object measured (e.g., '2kg' means the same weight everywhere; criterion 2). This differs from prices, customer ratings and other quantitative values that are attributed to some objects (e.g., apples) for some purposes and uses (e.g., trade, advertising). These pragmatic quantifications depend on considerations that go beyond the objects' specific properties and therefore vary across contexts and times, as does their specific quantitative meaning.

Quantitative psychologists' 'measurement' jargon alludes to the *epistemic authority* of genuine measurement. This misleads the public (Barrett, 2003, 2018). It also leads researchers themselves to mistake their purely pragmatic research frameworks for the realist framework required for measurement, thereby misguiding concepts and theories.

Psychology's pragmatic quantifications (e.g., rating data, IQ scores) and statistical analyses (e.g., psychometrics) are useful for distinguishing individuals by their observable responses or performances as well as for making decisions and predictions on the basis of the differences and relations observed. But these approaches do not constitute measurement because they neither conceptualise nor empirically implement unbroken traceable connections between the results and the quantities to be measured (measurands) in the actual study phenomena. By adapting the results instead to statistically useful data structures (e.g., group differences), these approaches cannot explore the performances or responses observed for their underlying causes. These result-dependent methods thus preclude explorations of the *actual study phenomena*, such as what specific intellectual abilities individuals may use to solve a task or what they consider in their ratings.

In sum, psychology must address the *gap* that often exist between its numerical data and statistical models, on the one side, and its actual study phenomena and the specific quantities to be measured in them (measurands), on the other. To bridge this gap, it must advance *genuine analogues* of measurement.

### 4.2 Genuine analogues of measurement: Elaborating quantitative psychology's epistemological and methodological fundamentals

Rosen's process model conceptualises the system of interrelated modelling relations, which is generally necessary to develop formal models that are appropriate for exploring real study systems in empirical sciences (Figure 1). Psychology's challenge lies in the necessity to advance for this general process model specific concepts and practices that meet the peculiarities of its study phenomena and language-based methods. This is because quantitative analysis can be informative only when the system of modelling relations is also *empirically implemented*—both semantically and syntactically—rather than just presumed on the basis of the methods' *inbuilt semantics* and researchers' own syntactic assignments—that quantitative analysis can be informative at all.

#### 4.2.1 Metrological frameworks: Adaptations to psychological research are appropriate only on the more abstract philosophy-of-science level

Metrology enables accurate and precise measurement of quantities in non-living phenomena featuring invariant (unchangeable) relations. Interdisciplinary attempts to translate and apply metrological concepts rather directly to psychology (esp. psychometrics), however, overlook fundamental ontic differences in its complex study phenomena. These involve, amongst others, variable and context-dependent relations (e.g., many–to–one, many–to–many), novel emergent properties and dynamic multi-level feedback loops leading to continuous change and development of parts and wholes. Therefore, specific

relations from observable phenomena (e.g., specific behaviours or test performances) to non-observable ones (e.g., specific intentions or intellectual abilities) that apply *to all individuals in all contexts and all times*—thus, that are invariant (one–to–one)—cannot be presumed. The study phenomena's *non-ergodicity* (non-equal synchronic and diachronic variations), as well, invalidates inferences from sample-level averages to measurands in single individuals.

Moreover, unlike metrology, psychology explores not just observable phenomena and their possibly underlying causes in themselves but also, and in particular, individuals' subjective and inter-subjective explanations, interpretations and appraisals of them. These multi-referential objects of research are conceptualised as *constructs* and require language-based methods for their exploration (Uher, 2022b, 2023a,b). Personality ratings, for example, were shown to be influenced by raters' knowledge of the phenomena and persons to be judged, raters' attitudes and relationships to them as well as raters' interpretation and use of the 'scales' (e.g., items' inbuilt semantics, redundancy, social valences), leading to guessing, inattention and bias (e.g., centrality tendency, social desirability, stereotyping, halo effect; Kenny, 1994; Leising et al., 2025; Podsakoff et al., 2003; Shweder and D'Andrade, 1980; Tourangeau et al., 2000; Uher and Visalberghi, 2016; Uher et al., 2013b). That is, raters interact differently with the same 'instrument', and even if they consider the same study phenomena in the same persons, they may invoke, in their ratings, different interpretational perspectives on them as well as indefinitely complex contexts. All this entails that rating data represent far more than just an observable 'reality' and always reflect various strong influences *apart* from that concrete 'reality' as well (Leising and Schilling, 2025).

That is, both psychology's complex study phenomena and its language-based 'instruments' are *rich in interpretable information*. In metrological frameworks, by contrast, information is conceptualised only as the outcome of measurement, in the formal model, whereas the real study system comprises the physical objects studied, those used as instruments as well as their empirical interaction (Figure 2; Mari et al., 2021). Therefore, metrological concepts cannot account for different interpretive perspectives that persons (raters and researchers alike) can flexibly and intentionally take on the same object of research as well as on the same 'instrument' and which are described with psychological constructs. Their conceptualisation is of no interest to metrology and physics but essential for psychology.

Still, as this article demonstrates, psychology can capitalise on metrology's theoretical fundamentals —just on far more abstract levels than interdisciplinary approaches can consider. This requires *transdisciplinary approaches*, as used here, to first make explicit and compare the different disciplines' basic ontological and epistemological presuppositions. This was a prerequisite for identifying the two abstract methodological principles (e.g., data generation traceability and numerical traceability) that implicitly underlie the metrological framework as well as for highlighting its direct conceptual connections to Rosen's general process scheme. The abstract philosophy of science perspective taken in transdisciplinarity is also essential to elaborate the ways in which concepts of physics and metrology, such as the problems of measurement and measurement coordination, can be meaningfully adapted to psychology to develop *genuine analogues of measurement* that are appropriate for its study phenomena's peculiarities (Uher, 2018a, 2019, 2020b, 2022a,b, 2023a, 2024).

### 4.2.2 Epistemically justified evidence for psychological research and applied practice: Requirements and challenges

Researchers and practitioners in applied settings increasingly highlight that testing theories, hypotheses and the effectiveness of interventions as well as making decisions about individuals, such as in clinical, educational and legal settings, require epistemically justified evidence of the phenomena studied—which the result-dependent approaches of rating methods and psychometrics cannot provide (Barrett, 2003, 2018; Faust, 2012; Heine and Heene, 2024; Hobart et al., 2007; Mislevy, 2024; Rosenbaum and Valsiner, 2011; Truijens, 2017; Uher, 2022b, 2023a). It is therefore crucial to *remedy the empirical breaks* that often occur between psychology's study phenomena and its pertinent data and models (Figure 7). This requires elaborate concepts and approaches of *scientific representation* that allow for establishing unbroken traceable connections that are appropriate for mapping formal systems onto the peculiarities of psychology's study phenomena (arrow 2, Figure 2). To achieve this, psychology must also *advance its ontological and epistemological fundamentals* (Fahrenberg, 2013, 2015; Hartmann, 1964; Lundh, 2018; Poli, 2006; Uher, 2021b). It must also advance its *methodology*, such as to specify the abilities that data generation methods must have for capturing specific properties in the study phenomena and for establishing traceable relations with them (Uher, 2013, 2015c, 2018a; Valsiner, 2017).

All these fundamentals are underdeveloped in quantitative psychology. Much of its numerical data are still generated with a simple yet seriously flawed method, developed already a century ago but still lacking a conceptual foundation. The common belief that rating 'scales' could enable standardised quantitative inquiries, implying that all individuals respond to standardised 'verbal stimuli' in the same ways and produce 'instrument' indications that allow for making straightforward inferences on the phenomena described, is unwarranted. It is surprising—if not paradoxical—that psychometricians claim that rating 'scales' enable the 'measurement' of individual variations while ignoring, at the same time, pronounced individual variations in the interpretation and use of these very same 'scales'. Psychology's challenges arise from the peculiarities of its study phenomena (e.g., higher-order complexity, non-ergodicity) and language-based methods (e.g., inbuilt semantics). These make it impossible to establish coherent measurement models that enable inferences from standardised instrument indications to non-observable measurands that could be reliable and valid *for all individuals in all contexts and times*. That is, *psychology's problems of measurement, measurement coordination and calibration cannot be solved on the sample level.*

Meanwhile, psychology as a science in general is more advanced and acknowledges that researchers' own assumptions, beliefs, thinking and judgements can (unintentionally) influence their research methods, theories and findings (Danziger, 1997; Fahrenberg, 2013; Fleck, 1935; James, 1890; Marsico et al., 2015;

Uher, 2013, 2015b; Weber, 1949). Quantitative psychology is still lacking behind these advancements (but see Jamieson et al., 2023). The common belief that quantitative methods could be generally more objective and free of subjectivity ('scientific')—and thus, superior to others *per se* (*quantificationism*)—is erroneous (Strauch, 1976; Uher, 2022b). Quantitative psychology must acknowledge the fact that, given the peculiarities of its study phenomena and of the language-based methods required for their investigation, (lay) persons (e.g., participants, patients) play a crucial epistemic role in the data generation process. As an empirical science, psychology cannot build just on the researchers' own inferences from the inbuilt semantics of their language-based methods and on their own assignments of syntactic structures to their data and models. Such practices are prone to ethnocentric and egocentric biases on the researchers' part, leading to distorted theories and findings (Uher, 2015b, 2020a).

To justify the use of rating 'scales' in psychological research and practice, it is of foremost importance to conceptualise and empirically explore how raters actually interpret and use these 'instruments'. This is a prerequisite for establishing traceable, coordinated and calibrated connections from the study phenomena and known reference quantities to the generated results (data generation traceability, numerical traceability)—thus, for establishing genuine analogues of measurement (Figure 2; Uher, 2018a, 2019, 2022b, 2023a).

### 4.2.3 Tackling psychology's problems of measurement coordination and calibration on the individual level: Empirical examples

Various lines of clinical research (e.g., on quality of life, chronic disease and therapeutic efficacy) already explored these problems under terms such as self-rated health (Fayers and Sprangers, 2002), patient-reported outcomes (PRO; Schwartz and Rapkin, 2004) and patient-centred measurement (PCM; Howard et al., 2024; McClimans, 2024; Pesudovs, 2006). They explicitly consider the fact that patients not only experience different symptoms, to different degrees and in different ways but also have diverse and changeable perspectives of their disease, treatment and quality of life. These researches consider that such complex study phenomena require for their description multi-referential conceptual systems (constructs) and language-based methods for their empirical investigation. Accordingly, they conceptualise in their methodological fundamentals the fact that patients' self-ratings involve perceptions, judgements, appraisals and also idiosyncratic criteria (Bosdet et al., 2021; Carr and Higginson, 2001; Kazdin, 2006; Schwartz and Rapkin, 2004; Truijens et al., 2019b).

This explicit conceptualisation is crucial to explain the frequent finding that changes in patients' self-ratings (e.g., pre-post treatment) often cannot be fully explained by actual changes in their health problems. Such *response shifts* pose challenges for evidence-based evaluations of clinical theories, treatments and therapies. They also question the utility of psychometric approaches for establishing the reliability and validity of assessment 'scales'. Response shifts were shown to occur for various reasons. First, they arise from patients' context-specific local interpretation of rating 'scales'. Furthermore, patients' interactions with the verbal descriptions of their symptoms on the 'instruments' can change how they interpret their symptoms, how they understand and experience their own condition and thus, the meaning that these have for them. Response shifts may also be due to changes in patients' subjective frames of reference, the standards of comparison that they consider, the relative importance that they ascribe to symptoms, their recall and sampling of salient experiences, how they combine their appraisals when choosing an answer box on the 'scale', and others (Desmet et al., 2021; Schwartz and Rapkin, 2004; Truijens et al., 2022; Vanier et al., 2021).

These findings illustrate why breaks in data generation traceability and numerical traceability occur when rating data are interpreted solely on the basis of their *inbuilt semantics* and *researcher-assigned syntax* (Figure 7). These and other lines of research demonstrated that raters' complex meaning-making processes must be considered to establish the *empirical semantics and syntax* of rating data—thus, their *epistemic validity*. Epistemic dialogue and other participative approaches involve both raters' first-person perspective and researchers' second-person perspective in order to probe into researchers' interpretation of raters' responses on standardised 'scales'. This allows for establishing feedback loops between the real study system and its formal model (e.g., data) in order to coordinate and calibrate their *empirical semantic and syntactic relations* (Lahlou et al., 2015; McClimans, 2024; Truijens et al., 2019a; Uher, 2018a, 2022b, 2023a). These lines of research show that *psychology's problems of measurement and of measurement coordination can be tackled on the individual level.*

### 4.2.4 Establishing the data's empirical semantics and syntax: Textual data from individuals' unrestricted verbal expressions vs. standardised rating data

To tackle these problems and to establish the data's epistemic validity, psychology must advance efficient methods for studying verbal descriptions that the studied *individuals themselves* find most appropriate to express their experiences. As George Kelly stated

> "… each person seeks to communicate his [her] distress in the terms that make sense to him [her], but not necessarily in terms that make sense to others" (Kelly, 1969, p. 58).

This requires methods for recording individuals' experiences and perspectives *without restricting their possibilities to verbally express themselves*. This insight is essential for conceptualising psychology's measurement problem. Specifically, individuals' interactions with a language-based 'instrument' (e.g., survey question) and the phenomena under study (e.g., anxieties) as well as individuals' indications of the outcomes of these interactions must be *unrestricted* and *adaptable*. Such methods allow for establishing relations in the real study system that are *meaningful for these individuals themselves*. This is crucial for making the observable (verbal) indications that raters produce informative about the—for researchers—non-observable study phenomena and their occurrences to which only raters have access (arrow 1, Figure 9). This methodical requirement arises from the complex relations (e.g., many–to–one) in psychology's study phenomena.

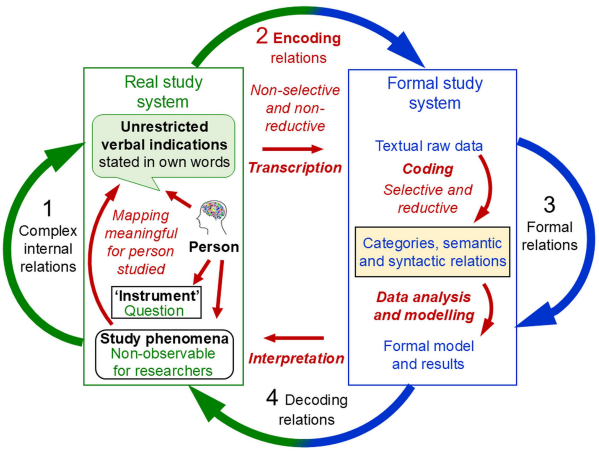



FIGURE 9
Establishing coherent systems of interrelated modelling relations through traceable encoding and analysis of individuals' unrestricted verbal expressions and ex post facto categorisation of study phenomena.

Metrological measurement models, by contrast, can deal only with unchangeable one–to–one relations of non-living nature, which can be identified through identically repeatable experiments. For this reason, the problems of coordination and calibration can be tackled on the sample level in metrology.

Individuals' indications, expressed in their own words, can be transcribed (e.g., verbatim) into textual data (or obtained from them in writing). This establishes documented, traceable and contextualised—yet *non-selective and non-reductive—encoding relations* between real and formal study system (arrow 2, Figure 9). The thus-generated textual raw data are then coded, whereby elements of individuals' encoded verbal statements are categorised into variables for further analysis. This establishes *selective and reductive coding relations*, which are likewise contextualised, unbroken, documented and traceable.[18] Thus, crucially, the *selective reductive mapping of the real system's open domain to the closed sign system used as its model does not occur in the encoding relations between real and formal system* (arrow 2), as conceptualised in Rosen's system. Instead, it occurs in an *additional coding relation within the formal study system* (arrow 3). This additional step of formal analysis accounts for the study phenomena's complexity, which makes attempts for *a priori* or *ad hoc* selective reduction prone to reductionist biases on the researchers' part.

Methods of text analysis (e.g., data mining; content, thematic or discourse analysis[19]) provide strategies to systematically analyse textual data, such as for specific words, word sequences or word proximities but also for specific contents, recurrent themes, concepts or discursive elements, often coded in fuzzy categories. These can also be further analysed for their occurrences (e.g., frequencies, associations and configurations)—thus, for syntactic (e.g., quantitative) relations. *Transparency in the selection and reduction decisions* during coding and analysis makes the formal model and the results thus-derived as well as their quantitative meanings traceable to concrete occurrences of verbally described events. By implementing data generation traceability and numerical traceability through iterative coordination and calibration processes, the model's empirical semantics and syntax are established—thus, *genuine* analogues of measurement (Uher, 2022a,b, 2023a).

The known challenges of some of these text analyses (e.g., coding biases, limited generalisability) testify to the complexity of the analytical and interpretational decisions, which are always required to scientifically categorise—thus, to selectively reduce and semiotically represent—psychology's complex study phenomena and to identify meaningful syntactic relations in them. These challenges become directly apparent because, in these methods, they are dealt with in the *formal study system*, where they can be explored in *documented traceable* ways *by the researchers themselves* (arrow 3, Figure 9). This also means that information about the study phenomena, as verbally described by the individuals experiencing them (arrow 1) and textually encoded in the formal study system (arrow 2), is scientifically categorised *ex post facto*—after the events to be studied have occurred in the real system. This is essential because, in complex and context-dependent phenomena, it cannot be predicted which specific events may occur. For this reason, data generated with open-ended response formats or participatory procedures can provide rich and in-depth insights into human experience, as clinical research has demonstrated (e.g., on response shifts).

Conceptualising the measurement problem for rating methods, by contrast, reveals a very different process. For ratings, researchers categorise their study phenomena aligned to their research questions and own preconceived ideas *ex ante*—before knowing which specific events of interest may actually occur in the real system studied (e.g., individuals). Researchers verbally describe these categories in statements whose general meaning derives from their *inbuilt semantics*—because no specific events to be described have yet occurred. These *ex ante* categorisations, which also serve as standardised 'instrument' indications (e.g., items), therefore need not be meaningful or even relevant to describe raters' concrete experiences and perspectives. Left without other options for expressing themselves, raters must adapt their interactions and judgements to the rating 'scale' provided, thereby producing indications that are less informative, if at all, about the study phenomena and raters' views on them. This entails several breaks in traceability (Figure 7).

18 These are specified, for example, in the methodological and epistemological foundations of a given method as well as in internationally agreed reporting standards (e.g., for so-called 'qualitative' methods in Levitt et al., 2018).

19 Some of these methods are commonly called 'qualitative', as opposed to 'quantitative' ones. This polarisation overlooks, however, that any quantity is always *of something*—a specific quality (Kaplan, 1964). Quantities are divisible properties of entities of the same kind—the same quality. Anything that is to be quantified must first be qualified in terms of the kind of entity that it is (Hartmann, 1964). Moreover, many so-called 'qualitative' methods establish data generation traceability and numerical traceability, thus meet the epistemic criteria of measurement, whereas rating methods, commonly regarded as 'quantitative', do not (Uher, 2022a,b, 2023a).

Yet, these breaks do not become apparent because the intricate decisions of how to relate the study phenomena's structures and occurrences to the fixed 'instrument' indications, both semantically and syntactically, are left to raters' intuitive decisions. Raters construe local meanings for standardised 'scales' to make them meaningful for their specific experiences and contexts. But how specifically the single raters interact with the methods ('instruments') and the study phenomena and thus, in what ways their observable indications can provide epistemically justifiable information about these phenomena remains *undocumented* and *non-traceable*. These relations are complex, variable, context-dependent and changeable. Therefore, they cannot be studied experimentally (unlike the one–to–one relations studied in metrology). Thus, in ratings, the *selective reductive mapping of the study phenomena's open domain to a closed sign system already occurs in the real study system*, inaccessible to researchers (arrow 1, Figure 10). This masks the tremendous challenges involved in the selective reduction of psychology's study phenomena. Moreover, this closed sign system itself (e.g., item statements) is aligned not to the specific events to be studied, as these have not yet occurred (*ex ante*), but to researchers' own preconceived ideas and study questions.

Researchers then encode raters' chosen indications using isomorphic mapping relations into rating data (arrow 2, Figure 10). Each standardised item statement is mapped to one item variable and interpreted regarding the general meaning of its *inbuilt semantics*. Raters' chosen answer boxes are rigidly encoded into predefined numerical values to which researchers attribute a desired syntax (e.g., quantitative meaning). As we have seen, this operationalist procedure introduces further breaks in the *empirical semantic and syntactic relations* between the rating data and the actual study phenomena (Figure 7). But these breaks often go unnoticed because, for rating methods, reporting standards demand traceability (transparency) only for the research design and statistical analyses (Appelbaum et al., 2018). But they do not also demand the data variables and values to be traceable back to occurrences of the study phenomena (as required, e.g., for ethological observations or software-based coding of behaviour). Therefore, rating methods preclude the conceptualisation and empirical implementation of coherent systems of interrelated modelling relations—and thus, of genuine analogues of measurement (Uher, 2018a, 2022b, 2023a).

In sum, psychology must invest more efforts to establish the epistemic validity of its data and models. These efforts can benefit from the advances made in artificial intelligence.

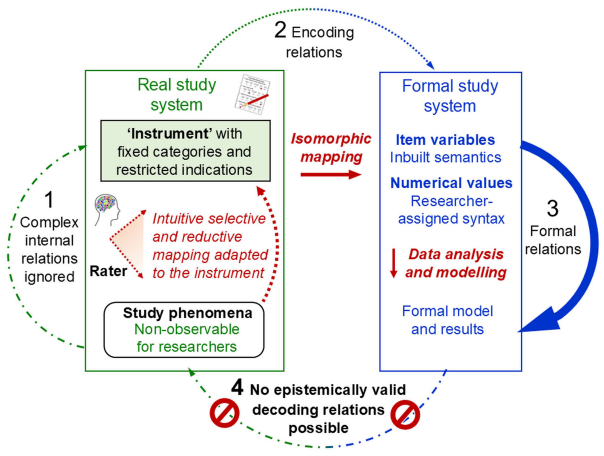



FIGURE 10
Rating methods: Ex ante categorisation of study phenomena and restricted instrument indications leave the crucial selective reduction decisions to raters' undocumented and non-traceable considerations.

## 4.3 Artificial intelligence: Language algorithms can support psychological research but also perpetuate psychologists' cardinal error

Psychology's language-based research can capitalise on the powerful artificial intelligence (AI) technologies that are modelling human language and that are now available at large scale—especially NLP algorithms (Section 3.6.4) and *large language models (LLMs)*. These deep learning machines capitalise on the foundations of NLPs but build their own internal implicit algorithms from processing vast textual data sets (e.g., books and websites). This extensive training enables LLMs to identify, predict and generate patterns and relations in human languages with higher adaptability, coherence and contextual relevance than previous NLPs. Therefore, they can 'understand' complex context, generate human-quality texts with human-like fluency and 'converse' in human-like fashion (e.g., ChatGPT).

These performances can meaningfully support psychological research. But they also trigger our deep-rooted natural tendency to attribute human characteristics to non-human entities (Hume, 1757). We focus on what appears to be human-like—that is, *anthropo-morphic*—but tend to ignore what is human-unlike (anthropo-centric biases type I and II (Uher, 2015b, 2020a). This anthropo-centrism profoundly shapes also how we perceive and engage with AI machines, thereby misleading our understanding of their capabilities and limitations (Yildiz, 2025). This applies in particular to the challenges and pitfalls inherent to language-based AI machines—especially those arising from their inbuilt semantics.

### 4.3.1 Efficient transcription and analysis of individuals' local context-specific meanings expressed in their own words through NLP algorithms and LLMs

Language algorithms can be used to efficiently analyse individuals' unrestricted verbal expressions—from transcription to the extraction of semantic and syntactic relations in documented traceable ways. Clinical researchers again pioneered in advancing methods for capturing and analysing the complexity of individuals' health conditions. They showed how patients' responses to well-prompted open-ended questions, expressed in their own words, can be analysed using machine learning techniques of NLP algorithms and LLMs. Their enhanced capabilities for analysing language

context enabled more detailed and more accurate assessments of patients' heterogeneous and complex mental health conditions than psychometric 'scales'—while also being individualised and efficient. Algorithm-based categorisations of open-ended self-descriptions discriminated even better between persons diagnosed with specific clinical conditions (e.g., anxiety, depression) and healthy persons than did pertinent self-ratings—although psychometric 'scales' are statistically designed and selected for enabling such discriminations reliably (Islam and Layek, 2023; Kerz et al., 2023; Kjell K. et al., 2021; Kjell O. et al., 2021; Sikström et al., 2023; Tabesh et al., 2025).

Hence, NLP algorithms and LLMs can be used to efficiently analyse individuals' local context-specific meanings, expressed in their own words, and to extract, summarise and categorise their general meanings using the AI models' *inbuilt semantics*. Their algorithmic parameters can also extract *syntactical* (e.g., quantitative) information (e.g., frequencies, associations) to enable further analysis of the identified categories (e.g., group comparisons). This procedure implements documented and traceable modelling relations between individuals' verbally described experiences (real study system) and the coded data and models about them (formal study system). This allows for establishing the results' *empirical semantics and syntax*—thus, their epistemic validity as required for genuine analogues of measurement. Proponents of rating methods, by contrast, still adhere to the inverse—yet epistemically invalid—procedure and therefore use language algorithms for other purposes.

### 4.3.2 Designing rating 'scales' with language algorithms cannot establish the data's empirical semantics and syntax as needed for genuine measurement analogues

Quantitative psychologists increasingly use NLPs and LLMs to design or improve psychometric 'scales'. Some aim to reduce the semantic overlap between 'scales' (Huang et al., 2025), to improve the content validity for specific constructs (Hernandez and Nie, 2023) or to tackle the incommensurability of constructs and operationalisations across studies (Wulff and Mata, 2023). Others aim to improve the prediction of human interpretation for more "robust, objective assessments" and to "enhance the scientific rigour" of psychometric tests (Milano et al., 2025). Thus, the *inbuilt semantics* of language algorithms is used to predefine categorisations of study phenomena (standardised statements). Their general meanings then serve as both 'instruments' and item variables to explore individuals' local context-specific meanings of their experiences and views on them. But as we have seen, these result-dependent procedures lead to several breaks in the thus-generated data's and models' traceability back to the phenomena studied in the real system (Figure 7). That is, they fail to establish the resulting model's *empirical semantics and syntax*—its *epistemic validity*.

This increasingly popular approach corresponds to creating a city map using well-established cartographic symbols and structures (e.g., for roads, buildings) yet *without mapping it empirically* onto a real city. It creates not a *map* but just an image of a city that may but need not exist as depicted. This is also like polishing the food descriptions on a restaurant's menu on the basis of what can generally be cooked, regardless of what dishes are actually cooked on a given day. Using AI algorithms of human languages to design psychometric 'instruments' cannot remedy the empirical breaks between real and formal study system.

Moreover, the basic idea is not new. Lexical approaches in differential psychology capitalise on the inbuilt semantics of natural languages, building on the assumption that those individual differences that are most salient will eventually become encoded in words. This *lexical hypothesis* (Galton, 1884; Klages, 1926) provided a stringent rationale for using the person-descriptive words in our natural languages to identify a few major dimensions of individual differences that are considered most important in folk psychology. This rationale underlies many popular 'personality' models developed over the last century (e.g., Big Five, 16PF; HEXACO; Allport and Odbert, 1936; John et al., 1988; Uher, 2013, 2015c, 2018b).

Despite its enormous importance for taxonomic research, however, the lexical hypothesis itself remained untested—even 141 years after its first articulation (Toomela, 2010; Uher, 2013; Westen, 1996). Still little is known about what specifically gets encoded in a language and how, what may be missed out and why. Humans invented an estimated 31,000 languages, of which only some 7,000 still exist (Crystal, 2000). Their vocabularies differ in what they allow us to describe. Their rules are extremely diverse, involving not just different scripts and speech patterns (signifiers) but also different rules that enable and enforce the communication of different types of information. In different languages, for example, communicators *either cannot or must indicate*—such as by modifying word endings—the reference to time (tense) and/or the extension over time (aspect); the agent (voice), state of completion and/or intentionality of actions; the grammatical gender and/or number of persons, objects, their attributes and/or actions; the syntactic function of persons, objects and events in a sentence (declension); the communicator's relation to the recipient, intention for communicating and/or source of the information communicated (e.g., whether from own observation, hearsay and/or inference), and others (Boroditsky, 2018; Deutscher, 2006, 2010).

That is, everyday language encodes everyday knowledge with all its socio-cultural biases and insufficiencies. If the everyday knowledges encoded in the semantics, syntaxes and pragmatics of our natural languages were epistemically valid and sufficiently accurate to describe and explain the structures and functions of human psyche, behaviour and society, then language scientists (e.g., linguists, philologists) would have long accomplished this task. But given the tremendous differences between languages, this strategy is epistemically not justified. Indeed, most AI technologies were developed in English. English is a mongrel language whose grammar was simplified already during the Mediaeval ages, when it was synthesised from Old English, Welsh, Gaelic, Danish, Norse, French, Old German and other languages. A focus on English-language algorithms will inevitably introduce ethno-centric biases, as happened before when Anglo-American 'personality' models (e.g., Five Factor Model) were claimed to be 'universally' valid for all human cultures (Uher, 2015c, 2018b).

Language algorithms are trained to identify and *re*-produce structures in human language—that is, they are modelling *human-produced* text or speech. But they cannot and do not establish relations (meanings) from the written or spoken sentences (signifiers) to the real 'world' (referents) that is being described in the language they are modelling. It is us, as humans, who construe, in our minds, these semantic relations to the real 'world' described (meanings). Meanings decay with individuals' minds (e.g., in dementia) and with their lives (Uher, 2015a). Therefore, languages die out with the persons using them (Crystal, 2000).

Language-based algorithms merely *re*-produce signifiers (words) and structures between them in ways that correspond to those that we use our languages. These structures were created through the efforts of past generations to mentally and semiotically represent the real 'world' around us and to communicate about it. AI systems meanwhile mimic these human-built structures in such sophisticated ways that we can easily integrate them into our thinking. This makes us inclined to attribute to the machine our own thinking of the semantic relations, which are built into our language and internalised in our minds. But we tend to overlook the fact that it is us who are thinking these relations, not the machine. This becomes obvious when we look at texts generated in a language foreign to us. Without having internalised its semantics, we cannot make sense of what is written—not mentally relate it to what it stands for in the real 'world' described. The machine cannot do this for us.

Our human abilities to immediately and effortlessly relate our language to the real 'world' described often leads us to overlook the crucial difference between the study phenomena and the means of their exploration (e.g., descriptions). To avoid confusing ontological and epistemological concepts—psychologists' cardinal error—psychologists should have at least some basic knowledge of human language. This is also necessary to use language-based algorithms in epistemically justified ways to advance psychological research.

One of psychology's key challenges lies in the fact that it must necessarily rely on lay people's abilities and their everyday language. This requires engaging with the individuals studied rather than distancing ourselves ever more from them by studying just standardised abstract descriptions of collective meanings that are predefined by researchers or AI machines. A science of psychology should advance approaches and methods that are epistemically justified for exploring its study phenomena in the specifics and contexts of their occurrences. Therefore, we need to know how individuals use their natural language and relate it to the real 'world' as they experience and see it in their given contexts. This knowledge will be crucial to systematically connect psychology's language-based data and formal models with the real-world phenomena that these are meant to represent and for which they serve only as surrogates—thus, to establish *genuine* analogues of measurement.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the [patients/participants OR patients/participants legal guardian/next of kin] was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## Author contributions

JU: Conceptualization, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

## Publisher's note

# References


Abel, D. L. (2009). The capabilities of chaos and complexity. *Int. J. Mol. Sci.* 10, 247–291. doi: 10.3390/ijms10010247

Abel, D. L. (2012). Is life unique? *Life* 2, 106–134. doi: 10.3390/life2010106

Abran, A., Desharnais, J.-M., and Cuadrado-Gallego, J. J. (2012). Measurement and quantification are not the same: ISO 15939 and ISO 9126. *J. Softw.* 24, 585–601. doi: 10.1002/smr.496

Allevard, T., Benoit, E., and Foulloy, L. (2005). Dynamic gesture recognition using signal processing based on fuzzy nominal scales. *Measurement* 38, 303–312. doi: 10.1016/j.measurement.2005.09.007

Allport, G. W., and Odbert, H. S. (1936). Trait names: a psycholexical study. *Psychol. Monogr.* 47, 1–171. doi: 10.1037/h0093360

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., and Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *Am. Psychol.* 73, 3–25. doi: 10.1037/amp0000191

Arnulf, J. K. (2020). "Wittgenstein's revenge: how semantic algorithms can help survey research escape Smedslund's labyrinth," in *Respect for thought*, eds. T. G. Lindstad, E. Stänicke, and J. Valsiner (Cham: Springer International Publishing), 285–307. doi: 10.1007/978-3-030-43066-5_17

Arnulf, J. K., and Larsen, K. R. (2015). Overlapping semantics of leadership and heroism: Expectations of omnipotence, identification with ideal leaders and disappointment in real managers. *Scand. Psychol.* 2:e3. Available online at: https://ssrn.com/abstract=3716109 (accessed July 30, 2023).

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018). Semantic algorithms can detect how media language shapes survey responses in organizational behaviour. *PLoS ONE* 13:e0207643. doi: 10.1371/journal.pone.0207643

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PLoS ONE* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Olsson, U. H., and Nimon, K. (2024). Measuring the menu, not the food: "psychometric" data may instead measure "lingometrics" (and miss its greatest potential). *Front. Psychol.* 15:1308098. doi: 10.3389/fpsyg.2024.1308098

Arro, G. (2013). Peeking into personality test answers: Inter- and intraindividual variety in item interpretations. *Integr. Psychol. Behav. Sci.* 47, 56–76. doi: 10.1007/s12124-012-9216-9

Atmanspacher, H. (1997). Cartesian cut, Heisenberg cut, and the concept of complexity. *World Fut.* 49, 333–355. doi: 10.1080/02604027.1997.9972639

Baianu, I. C., and Poli, R. (2011). From simple to highly complex systems: a paradigm shift towards non-Abelian emergent system dynamics and meta-levels. *Acta Univ. Apulen.* 26, 131–167. Available online at: https://www.emis.de/journals/AUA/pdf/57_410_no_10_acta_barni.pdf (accessed May 4, 2024).

Barrett, L. F., Mesquita, B., and Smith, E. R. (2010). "The context principle," in *The mind in context*, eds. B. Mesquita, L. F. Barrett, and E. R. Smith (New York, NY, US: The Guilford Press), 1–22.

Barrett, P. (2003). Beyond psychometrics. *J. Manag. Psychol.* 18, 421–439. doi: 10.1108/02683940310484026

Barrett, P. (2018). The EFPA test-review model: when good intentions meet a methodological thought disorder. *Behav. Sci.* 8:5. doi: 10.3390/bs8010005

Baumeister, R. F., Vohs, K. D., and Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: whatever happened to actual behavior? *Persp. Psychol. Sci.* 2, 396–403. doi: 10.1111/j.1745-6916.2007.00051.x

Bergman, L. R., and Trost, K. (2006). The person-oriented versus the variable-oriented approach: are they complementary, opposites, or exploring different worlds? *Merrill Palmer Q.* 52, 601–632. doi: 10.1353/mpq.2006.0023

Bernstein, J. H. (2015). Transdisciplinarity: a review of its origins, development, and current issues. *J. Res. Pract.* 11:412. https://jrp.icaap.org/index.php/jrp/article/view/510.html (accessed September 23, 2019).

BIPM (2019). *BIPM: The international system of units (SI) (9th ed).* Organisation Intergouvernementale de la Convention du Mètre. Available online at: https://www.bipm.org/documents/20126/41483022/SI-Brochure-9-EN.pdf (accessed October 24, 2024).

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proc. Nat. Acad. Sci.* 17, 656–660. doi: 10.1073/pnas.17.2.656

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic*. 36, 35–37.

Boroditsky, L. (2018). *7,000* Universes: *How the Languages We Speak Shape the Ways We Think*. Toronto: Doubleday Canada.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511490026

Borsboom, D. (2008). Latent variable theory. *Measurement* 6, 25–53. doi: 10.1080/15366360802035497

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., and Franić, S. (2009). "The end of construct validity," in *The concept of validity: Revisions, new directions, and applications* (Charlotte, NC: IAP Information Age Publishing), 135–170.

Borsboom, D., and Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory Psychol.* 14, 105–120. doi: 10.1177/0959354304040200

Bosdet, L., Herron, K., and Williams, A. C., de C. (2021). Exploration of hospital inpatients' use of the verbal rating scale of pain. *Front. Pain Res.* 2:723520. doi: 10.3389/fpain.2021.723520

Carr, A. J., and Higginson, I. J. (2001). Are quality of life measures patient centred? *BMJ.* 322, 1357–1360. doi: 10.1136/bmj.322.7298.1357

Chang, H. (2004). *Inventing* Temperature: *Measurement and Scientific Progress*. Oxford: Oxford University Press. doi: 10.1093/0195171276.001.0001

Cialdini, R. B. (2009). We have to break up. *Persp. Psychol. Sci.* 4, 5–6. doi: 10.1111/j.1745-6924.2009.01091.x

Cicchetti, D., and Rogosch, F. A. (1996). Equifinality and multifinality in developmental psychopathology. *Dev. Psychopathol.* 8, 597–600. doi: 10.1017/S0954579400007318

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Crystal, D. (2000). *Language Death*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139106856

Danziger, K. (1985). The methodological imperative in psychology. *Philos. Soc. Sci.* 15, 1–13. doi: 10.1177/004839318501500101

Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511524059

Danziger, K. (1997). *Naming the Mind: How Psychology Found its Language*. London, UK: Sage. doi: 10.4135/9781446221815

Danziger, K., and Dzinas, K. (1997). How psychology got its variables. *Canadian Psycholo.* 38, 43–48. doi: 10.1037/0708-5591.38.1.43

Dawes, R., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science* 243, 1668–1674. doi: 10.1126/science.2648573

Dennett, D. (2012). A perfect and beautiful machine': What Darwin's theory of evolution reveals about artificial intelligence. *The Atlantic*, *June 22*. Available online at: http://hdl.handle.net/10427/000489 (accessed October 23, 2024).

Desmet, M., Van Nieuwenhove, K., De Smet, M., Meganck, R., Deeren, B., Van Huele, I., et al. (2021). What too strict a method obscures about the validity of outcome measures. *Psychother. Res.* 31, 882–894. doi: 10.1080/10503307.2020.1865584

Deutscher, G. (2006). *The Unfolding of Language: The Evolution of Mankind's Greatest Invention*. London, UK: Arrow.

Deutscher, G. (2010). *Through the Language Glass: Why the World Looks Different in Other Languages.* New York, NY, US: Metropolitan Books/Henry Holt and Company.

Doliński, D. (2018). Is psychology still a science of behaviour? *Soc. Psychol. Bull.* 13:e25025. doi: 10.5964/spb.v13i2.25025

Eronen, M. I. (2024). Causal complexity and psychological measurement. *Philos. Psychol.* 1–16. doi: 10.1080/09515089.2023.2300693

Fahrenberg, J. (2013). *Zur Kategorienlehre der Psychologie: Komplementaritätsprinzip; Perspektiven und Perspektiven-Wechsel [On the category theory of psychology: Principle of complementarity, perspectives and changes of perspectives]*. Lengerich, Germany: Pabst Science Publishers.

Fahrenberg, J. (2015). *Theoretische Psychologie - Eine Systematik der Kontroversen*. Lengerich, Germany: Pabst Science Publishers.

Faust, D. (2012). *Ziskin's Coping With Psychiatric and Psychological Testimony*. Oxford: Oxford University Press. doi: 10.1093/med:psych/9780195174113.001.0001

Fayers, P. M., and Sprangers, M. A. G. (2002). Understanding self-rated health. *Lancet* 359, 187–188. doi: 10.1016/S0140-6736(02)07466-4

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement* 34, 39–48. doi: 10.1016/S0263-2241(03)00018-6

Fisher, W. P. (2009). Invariance and traceability for measures of human, social, and natural capital: theory and application. *Measurement* 42, 1278–1287. doi: 10.1016/j.measurement.2009.03.014

Fisher, W. P., and Pendrill, L. (2024). *Models,* Measurement*, and Metrology Extending the SI. Trust and Quality Assured Knowledge Infrastructures.* Berlin: De Gruyter Oldenbourg. doi: 10.1515/9783111036496

Fleck, L. (1935). *Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv*. Frankfurt am Main::

Suhrkamp. English translation: (1979). *The Genesis and Development of a Scientific Fact* (T. J. Trenn and R. K. Merton, Eds.) Chicago: University of Chicago Press.

Frigg, R., and Nguyen, J. (2021). "Scientific representation," in *The Stanford Encyclopedia of Philosophy* ed. E. Zalta. Available online at: https://plato.stanford.edu/archives/win2021/entries/scientific-representation (accessed June 23, 2023).

Galton, F. (1884). Measurement of character. *Fortnightly Rev.* 36, 179–185.

Gibbs, P., and Beavis, A. (2020). *Contemporary Thinking on Transdisciplinary Knowledge: What Those Who Know, Know*. Cham, Switzerland: Springer. doi: 10.1007/978-3-030-39785-2

Hance, J. R., and Hossenfelder, S. (2022). What does it take to solve the measurement problem? *J. Phys. Commun.* 6:102001. doi: 10.1088/2399-6528/ac96cf

Hartmann, N. (1964). *Der Aufbau der realen Welt. Grundriss der allgemeinen Kategorienlehre [The structure of the real world. Outline of the general theory of categories] 1940 (3rd ed.).* Berlin: Walter de Gruyter. doi: 10.1515/9783110823844

Harvard, S., and Winsberg, E. (2022). The epistemic risk in representation. *Kennedy Inst. Ethics J.* 31, 1–31. doi: 10.1353/ken.2022.0001

Heine, J.-H., and Heene, M. (2024). Measurement and mind: Unveiling the self-delusion of metrification in psychology. *Measurement* 2024, 1–29. doi: 10.1080/15366367.2024.2329958

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik [On The Actual Content of Quantum Theoretical Kinematics and Mechanics]. *Zeitschrift Für Physik*. 43, 172–198. doi: 10.1007/BF01397280

Hempel, C. G. (1952). "Fundamentals of concept formation in empirical science," in *International Encyclopedia of Unified Science,* ed. H. C. Gustav (Chicago: University of Chicago Press).

Hernandez, I., and Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Pers. Psychol.* 76, 1011–1035. doi: 10.1111/peps.12543

Hobart, J. C., Cano, S. J., Zajicek, J. P., and Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol.* 6, 1094–1105. doi: 10.1016/S1474-4422(07)70290-9

Howard, A. F., Warner, L., Cuthbertson, L., and Sawatzky, R. (2024). Patient-driven research priorities for patient-centered measurement. *BMC Health Serv. Res.* 24:735. doi: 10.1186/s12913-024-11182-x

Huang, Z., Long, Y., Peng, K., and Tong, S. (2025). *An embedding-based semantic analysis* approach: A preliminary study on redundancy detection in psychological concepts operationalized by scales. *J. Intell.* 13:11. doi: 10.3390/jintelligence13010011

Hume, D. (1757). *1957 The Natural History of Religion*. Stanford, CA: Stanford University Press.

Islam, R., and Layek, M. A. (2023). StackEnsembleMind: Enhancing well-being through accurate identification of human mental states using stack-based ensemble machine learning. *Inform. Med. Unlocked* 43:101405. doi: 10.1016/j.imu.2023.101405

James, W. (1890). *Principles of Psychology (Vol. 1)*. New York, NY: Holt. doi: 10.1037/10538-000

Jamieson, M. K., Govaart, G. H., and Pownall, M. (2023). Reflexivity in quantitative research: a rationale and beginner's guide. *Soc. Personal. Psychol. Compass* 17:e12735. doi: 10.1111/spc3.12735

JCGM100:2008 (2008). *Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Joint Committee for Guides in Metrology (originally published in 1993). doi: 10.59161/JCGM100-2008E

JCGM200:2012 (2012). *International vocabulary of metrology – Basic and general concepts and associated terms (VIM 3rd edition)*. Working Group 2 (Eds.), Joint Committee for Guides in Metrology. doi: 10.59161/JCGM200-2012

John, O. P., Angleitner, A., and Ostendorf, F. (1988). The lexical approach to personality: a historical review of trait taxonomic research. *Eur. J. Pers.* 2, 171–203. doi: 10.1002/per.2410020302

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Measur.* 50, 1–73. doi: 10.1111/jedm.12000

Kaplan, A. (1964). *The Conduct of Inquiry: Methodology for Behavioral Science*. Scranton, PA: Chandler Publishing Co.

Kazdin, A. E. (2006). Arbitrary metrics: implications for identifying evidence-based treatments. *Am. Psychol.* 61, 42–71. doi: 10.1037/0003-066X.61.1.42

Kellen, D., Davis-Stober, C. P., Dunn, J. C., and Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Persp. Psychol. Sci.* 16, 767–778. doi: 10.1177/1745691620974771

Kelly, G. (1955). *The Psychology of Personal Constructs (Volume 1 and 2)*. London, UK: Routledge.

Kelly, G. (1969). "The autobiography of a theory (1963)," in *Clinical psychology and personality: The selected papers of George Kelly,* ed. B. Maher (New York: Wiley), 46–65.

Kenny, D. A. (1994). *Interpersonal* Perception: *A Social Relations Analysis*. London: Guilford Press.

Kerz, E., Zanwar, S., Qiao, Y., and Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Front. Psychiat.* 14:1219479. doi: 10.3389/fpsyt.2023.1219479

Khatin-Zadeh, O., Eskandari, Z., Farsani, D., and Banaruee, H. (2025). Dynamic mathematical processing through symbolic, situational, and verbal representations. *Integr. Psychol. Behav. Sci.* 59:33. doi: 10.1007/s12124-025-09899-3

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* 82, 3713–3744. doi: 10.1007/s11042-022-13428-4

Kjell, K., Johnsson, P., and Sikström, S. (2021). Freely generated word responses analyzed with artificial intelligence predict self-reported symptoms of depression, anxiety, and worry. *Front. Psychol.* 12:602581. doi: 10.3389/fpsyg.2021.602581

Kjell, O., Daukantaite, D., and Sikström, S. (2021). Computational language assessments of harmony in life — Not satisfaction with life or rating scales — correlate with cooperative behaviors. *Front. Psychol.* 12:601679. doi: 10.3389/fpsyg.2021.601679

Klages, L. (1926). *Grundlagen der Charakterkunde [The science of character; W.H. Johnston, Trans. 1932]*. London: Allen and Unwin.

Korzybski, A. (1933). *Science and sanity. An introduction to non-Aristotelian systems and general semantics*. The International Non-Aristotelian Library Publication Company. Available online at: https://archive.org/details/sciencesanityint00korz/page/n7/mode/2up (accessed March 3, 2025).

Krantz, D., Luce, R. D., Tversky, A., and Suppes, P. (1971). *Foundations of Measurement Vol. I: Additive and Polynomial Representations*. San Diego: Academic Press. doi: 10.1016/B978-0-12-425401-5.50011-8

Lahlou, S., Le Bellu, S., and Boesen-Mariani, S. (2015). Subjective evidence based ethnography: method and applications. *Integr. Psychol. Behav. Sci.* 49, 216–238. doi: 10.1007/s12124-014-9288-9

Lakoff, G., and Johnsen, M. (2003). *Metaphors We Live by*. London: The University of Chicago press. doi: 10.7208/chicago/9780226470993.001.0001

Lamiell, J. (2018). From psychology to psychodemography: How the adoption of population-level statistical methods transformed psychological science. *Am. J. Psychol.* 131, 471–475. doi: 10.5406/amerjpsyc.131.4.0471

Lamiell, J. (2019). *Psychology's Misuse of Statistics and Persistent Dismissal of its Critics*. Cham: Palgrave Macmillan. doi: 10.1007/978-3-030-12131-0

Leinster, T. (2014). *Basic Category Theory*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781107360068

Leising, D., Borgstede, M., Burger, J., Zimmermann, J., Bäckström, M., Oltmanns, J., et al. (2025). Why do judgments on different person-descriptive attributes correlate with one another? A conceptual analysis with relevance for most psychometric research. *Collabra: Psychol.* 11:133683. doi: 10.1525/collabra.133683

Leising, D., and Schilling, R. L. (2025). A mathematical model of person judgment part 1: Cue emergence. *Person. Sci.* 6:27000710241291544. doi: 10.1177/27000710241291543

Lennox, J. (2024). *Robert Rosen and Relational System* Theory: *An Overview*. Cham: Springer. doi: 10.1007/978-3-031-51116-5

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., and Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. *Am. Psychol.* 73, 26–46. doi: 10.1037/amp0000151

Lewin, K. (1936). *Principles of Topological Psychology*. New York, NY: McGraw-Hill.

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 22, 1–55.

Linkov, V. (2024). Qualitative (pure) mathematics as an alternative to measurement. *Front. Psychol.* 15:1374308. doi: 10.3389/fpsyg.2024.1374308

Luce, R. D., Krantz, D., Suppes, P., and Tversky, A. (1990). *Foundations of Measurement, Vol. III: Representation, Axiomatization, and Invariance.* San Diego, CA, US. doi: 10.1016/B978-0-12-425403-9.50010-2

Luchetti, M. (2020). From successful measurement to the birth of a law: disentangling coordination in Ohm's scientific practice. *Stud. History Philos. Sci. Part A* 84, 119–131. doi: 10.1016/j.shpsa.2020.09.005

Luchetti, M. (2024). Epistemic circularity and measurement validity in quantitative psychology: insights from Fechner's psychophysics. *Front. Psychol.* 15:1354392. doi: 10.3389/fpsyg.2024.1354392

Lundh, L.-G. (2018). *Psychological science within a three-dimensional ontology. Integr. Psychol. Behav. Sci.* 52, 52–66. doi: 10.1007/s12124-017-9412-8

Lundmann, L., and Villadsen, J. W. (2016). Qualitative variations in personality inventories: subjective understandings of items in a personality inventory. *Qual. Res. Psychol.* 13, 166–187. doi: 10.1080/14780887.2015.1134737

Luria, A. (1966). *Higher Cortical Functions in Man*. New York: Basic Books.

Mach, E. (1986). *Principles of the Theory of Heat Historically and Critically Elucidated*. D. Reidel Publishing Company.

Maraun, M. D., and Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas Psychol.* 31, 32–42. doi: 10.1016/j.newideapsych.2011.02.006

Maraun, M. D., and Halpin, P. F. (2008). Manifest and latent variates. *Measurement* 6, 113–117. doi: 10.1080/15366360802035596

Margenau, H. (1950). *The Nature of Physical Reality. A Philosophy of Modern Physics*. New York: McGraw-Hill.

Mari, L., Carbone, P., Giordani, A., and Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Stud. History Philos. Sci.* 65, 46–56. doi: 10.1016/j.shpsa.2017.08.001

Mari, L., Carbone, P., and Petri, D. (2015). "Fundamentals of hard and soft measurement," in *Modern measurements: Fundamentals and applications,* eds. A. Ferrero, D. Petri, P. Carbone, and M. Catelani (Hoboken, NJ: John Wiley and Sons), 203–262. doi: 10.1002/9781119021315.ch7

Mari, L., Wilson, M., and Maul, A. (2021). *Measurement Across the Sciences. Developing a Shared Concept System For Measurement*. Cham: Springer. doi: 10.1007/978-3-030-65558-7

Marsico, G., Andrisano Ruggieri, R., and Salvatore, S. (2015). *Reflexivity and Psychology*. Charlotte, NC, USA: Information Age Publishing.

Mason, P. H. (2010). Degeneracy at multiple levels of complexity. *Biol. Theory* 5, 277–288. doi: 10.1162/BIOT_a_00041

McClimans, L. (2024). *Patient-Centered Measurement: Ethics,* Epistemology, *and Dialogue in Contemporary Medicine*. Oxford: Oxford University Press. doi: 10.1093/oso/9780197572078.001.0001

McClimans, L., Browne, J., and Cano, S. (2017). Clinical outcome measurement: Models, theory, psychometrics and practice. *Stud. History Philos. Sci.* 65, 67–73. doi: 10.1016/j.shpsa.2017.06.004

McGrane, J. A. (2015). Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Front. Psychol.* 6:431. doi: 10.3389/fpsyg.2015.00431

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychologist* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Michaelis, L. A. (2003). "Word meaning, sentence meaning, and syntactic meaning," in *Cognitive approaches to lexical semantics,* eds. H. Cuyckens, R. Dirven, and J. R. Taylor (Berlin, New York: De Gruyter Mouton), 163–210. doi: 10.1515/9783110219074.163

Michell, J. (1999). *Measurement in Psychology. A Critical History of a Methodological Concept*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511490040

Mikulecky, D. C. (2000). Robert Rosen: the well-posed question and its answer - why are organisms different from machines? *Syst. Res. Behav. Sci.* 17, 419–432. doi: 10.1002/1099-1743(200009/10)17:5<419::AID-SRES367>3.0.CO;2-D

Mikulecky, D. C. (2001). Robert Rosen (1934–1998): a snapshot of biology's Newton. *Comput. Chem.* 25, 317–327. doi: 10.1016/S0097-8485(01)00079-1

Mikulecky, D. C. (2011). Even more than life itself: beyond complexity. *Axiomathes* 21, 455–471. doi: 10.1007/s10516-010-9119-7

Milano, N., Luongo, M., Ponticorvo, M., and Marocco, D. (2025). Semantic analysis of test items through large language model embeddings predicts a-priori factorial structure of personality tests. *Curr. Res. Behav. Sci.* 8:100168. doi: 10.1016/j.crbeha.2025.100168

Mislevy, R. J. (2024). Sociocognitive and argumentation perspectives on psychometric modeling in educational assessment. *Psychometrika* 89, 64–83. doi: 10.1007/s11336-024-09966-5

Molenaar, P. C. M., and Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Curr. Dir. Psychol. Sci.* 18, 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Montuori, A. (2008). "Foreword: transdisciplinarity," in *Transdisciplinarity: Theory and practice,* ed. B. Nicolescu (Cresskill, NJ: Hampton Press).

Morin, E. (1992). *Method: Towards a Study of Humankind. Volume 1: The nature of nature.* (J. L. R. Belanger, Trans.). New York, US: Peter Lang.

Morin, E. (2008). *On Complexity*. Cresskill, NJ: Hampton Press.

Narens, L. (2002). A meaningful justification for the representational theory of measurement. *J. Math. Psychol.* 46, 746–768. doi: 10.1006/jmps.2002.1428

Newfield, C., Alexandrova, A., and John, S. (2022). *Limits of the Numerical: The Abuses and Uses of Quantification*. Chicago: University of Chicago Press. doi: 10.7208/chicago/9780226817163.001.0001

Nicolescu, B. (2002). *Manifesto of Transdisciplinarity 1996 (K.* C. Voss, Trans.). Albany, NY: State University of New York Press.

Nicolescu, B. (2008). "*In vitro* and *in vivo* knowledge: Methodology of transdisciplinarity," in *Transdisciplinarity: Theory and practice,* ed. B. Nicolescu (Cresskill, NJ: Hampton Press), 1–21.

Nowotny, H. (2005). The increase of complexity and its reduction: emergent interfaces between the natural sciences, humanities and social sciences. *Theory, Cult. Soc.* 22, 15–31. doi: 10.1177/0263276405057189

Ogden, C. K., and Richards, I. A. (1923). *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*. Orlando, FL: Harcourt, Brace and World.

Olsson, E. (2023). "Coherentist theories of epistemic justification," in *The Stanford Encyclopedia of Philosophy,* E. N. Zalta and U. Nodelman. Available online at: https://plato.stanford.edu/archives/win2023/entries/justep-coherence (accessed February 23, 2024).

Osborne-Crowley, K. (2020). Social cognition in the real world: Reconnecting the study of social cognition with social reality. *Rev. General Psychol.* 24, 144–158. doi: 10.1177/1089268020906483

Pattee, H. H. (2001). The physics of symbols: bridging the epistemic cut. *BioSystems* 60, 5–21. doi: 10.1016/S0303-2647(01)00104-6

Pattee, H. H. (2013). Epistemic, evolutionary, and physical conditions for biological information. *Biosemiotics* 6, 9–31. doi: 10.1007/s12304-012-9150-8

Pattee, H. H. (2021). Symbol grounding precedes interpretation. *Biosemiotics* 14, 561–568. doi: 10.1007/s12304-021-09458-4

Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce, Vols. 1-6, C. Hartshorne and P. Weiss (eds.), vols. 7-8, A. W. Burks (ed.).* Cambridge, MA: Harvard University Press.

Pesudovs, K. (2006). Patient-centred measurement in ophthalmology – a paradigm shift. *BMC Ophthalmol.* 6:25. doi: 10.1186/1471-2415-6-25

Piaget, J. (1972). "The epistemology of interdisciplinary relationships," in *Centre for Educational Research and Innovation (CERI), Interdisciplinarity: Problems of teaching and research in universities* (Paris, France: Organisation for Economic Co-operation and Development), 127–139.

Pirnay-Dummer, P., Ifenthaler, D., and Seel, N. M. (2012). "Semantic networks," in *Encyclopedia of the Sciences of Learning,* ed. N. M. Seel (Boston, MA: Springer US), 3025–3029. doi: 10.1007/978-1-4419-1428-6_1933

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879

Pohl, C. (2011). What is progress in transdisciplinary research? *Futures* 43, 618–626. doi: 10.1016/j.futures.2011.03.001

Poli, R. (2006). Levels of reality and the psychological stratum. *Rev. Int. Philos.* 236, 163–180. doi: 10.3917/rip.236.0163

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400821617

Ramage, M., and Shipp, K. (2020). *Systems Thinkers (2nd ed.)*. London, UK: Springer. doi: 10.1007/978-1-4471-7475-2

Rammstedt, B., and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J. Res. Pers.* 41, 203–212. doi: 10.1016/j.jrp.2006.02.001

Rashevsky, N. (1960a). Contributions to relational biology. *Bull. Math. Biophys.* 22, 73–84. doi: 10.1007/BF02477973

Rashevsky, N. (1960b). *Mathematical Biophysics: Physico-Mathematical Foundations of Biology (vol. 1–2), first published in 1938*. New York: Dover Publications.

Reichenbach, H. (1920). *Relativitätstheorie und Erkenntnis Apriori [Relativity theory and apriori knowledge]*. Berlin: Springer. doi: 10.1007/978-3-642-50774-8

Richters, J. E. (2021). Incredible utility: the lost causes and causal debris of psychological science. *Basic Appl. Soc. Psych.* 43, 366–405. doi: 10.1080/01973533.2021.1979003

Rød, J. (2004). Cartographic signs and arbitrariness. *Cartographica* 39, 27–36. doi: 10.3138/4462-2125-1312-217T

Romeijn, J.-W. (2017). "Philosophy of statistics," in *The Stanford Encyclopedia of Philosophy,* ed. E. Zalta. Available online at: https://plato.stanford.edu/archives/spr2017/entries/statistics (accessed January 21, 2024).

Rosen, R. (1970). *Dynamical Systems Theory in Biology*. New York: Wiley Interscience.

Rosen, R. (1985). *Anticipatory Systems: Philosophical,* Mathematical*, and Methodological Foundations.* New York: Elsevier Science and Technology Books.

Rosen, R. (1991). *Life itself. A comprehensive inquiry into the nature, origin, and fabrication of life*. New York: Columbia University Press.

Rosen, R. (1999). *Essays on Life itself*. New York: Columbia University Press.

Rosenbaum, P. J., and Valsiner, J. (2011). The un-making of a method: from rating scales to the study of psychological processes. *Theory Psychol.* 21, 47–65. doi: 10.1177/0959354309352913

Rudolph, L. (2013). *Qualitative Mathematics for the Social Sciences.* Mathematical Models for Research on Cultural Dynamics. (L. Rudolph, Ed.). London: Routledge. doi: 10.4324/9780203100806

Russell, Y. I. (2022). Three problems of interdisciplinarity. *Avant* 13, 1–19. doi: 10.26913/ava202206

Sato, T., Hidaka, T., and Fukuda, M. (2009). "Depicting the dynamics of living the life: The trajectory equifinality model," in *Dynamic process methodology in the social and developmental sciences,* eds. J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, and N. Chaudhary (New York, NY: Springer US), 217–240. doi: 10.1007/978-0-387-95922-1_10

Scholz, J. (2024). Agential realism as an alternative philosophy of science perspective for quantitative psychology. *Front. Psychol.* 15:1410047. doi: 10.3389/fpsyg.2024.1410047

Schrödinger, E. (1964). *What is life? Reprinted in: What is life? With Mind and Matter and Autobiographical Sketches*. Cambridge, UK: Cambridge University Press.

Schwager, K. W. (1991). The representational theory of measurement: an assessment. *Psychol. Bull.* 110, 618–626. doi: 10.1037/0033-2909.110.3.618

Schwartz, C. E., and Rapkin, B. D. (2004). Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health Qual. Life Outcomes* 2:16. doi: 10.1186/1477-7525-2-16

Sechrest, L., McKnight, P., and McKnight, K. (1996). *Calibration of measures for psychotherapy outcome studies. Am. Psychol.* 51, 1065–71. doi: 10.1037/0003-066X.51.10.1065

Shweder, R. A. (1977). Likeness and likelihood in everyday thought: magical thinking in judgments about personality. *Curr. Anthropol.* 18, 637–658. doi: 10.1086/201974

Shweder, R. A., and D'Andrade, R. G. (1980). "The systematic distortion hypothesis," in *Fallible judgment in behavioral research: New directions for methodology of social and behavioral science*, ed. R. A. Shweder (San Francisco: Jossey-Bass), 37–58.

Sikström, S., Pålsson Höök, A., and Kjell, O. (2023). Precise language responses versus easy rating scales—Comparing respondents' views with clinicians' belief of the respondent's views. *PLoS ONE* 18:e0267995. doi: 10.1371/journal.pone.0267995

Smedslund, G., Arnulf, J. K., and Smedslund, J. (2022). Is psychological science progressing? Explained variance in PsycINFO articles during the period 1956 to 2022. *Front. Psychol.* 13:1089089. doi: 10.3389/fpsyg.2022.1089089

Smedslund, J. (1991). The pseudoempirical in psychology and the case for psychologic. *Psychol. Inq.* 2, 325–338. doi: 10.1207/s15327965pli0204_1

Smedslund, J. (2004). *Dialogues About a New Psychology*. Chagrin Falls, OH: Taos Inst Publications.

Smedslund, J. (2016a). Practicing psychology without an empirical evidence-base: the bricoleur model. *New Ideas Psychol.* 43, 50–56. doi: 10.1016/j.newideapsych.2016.06.001

Smedslund, J. (2016b). Why psychology cannot be an empirical science. *Integr. Psychol. Behav. Sci.* 50, 185–195. doi: 10.1007/s12124-015-9339-x

Smedslund, J. (2021). From statistics to trust: psychology in transition. *New Ideas Psychol.* 61, 100848. doi: 10.1016/j.newideapsych.2020.100848

Speelman, C. P., and McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Front. Psychol.* 11:594675. doi: 10.3389/fpsyg.2020.594675

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 667–680. doi: 10.1126/science.103.2684.677

Stevens, S. S. (1957). On the psychophysical law. *Psychol. Rev.* 64, 153–181. doi: 10.1037/h0046162

Strauch, R. E. (1976). Critical look at quantitative methodology. *Policy Anal.* 2, 121–144.

Strom, D. J., and Tabatadze, G. (2022). Why "measurand" Is the first scientific word we should teach health physicists. *Health Phys.* 122, 607–613. doi: 10.1097/HP.0000000000001534

Suppes, P., Krantz, D., Luce, D., and Tversky, A. (1989). *Foundations of Measurement, Vol. II: Geometrical, Threshold, and Probabilistic Representations*. New York, NY: Academic Press. doi: 10.1016/B978-0-12-425402-2.50008-9

Suppes, P., and Zinnes, J. (1963). "*Basic measurement theory*," in *Handbook of Mathematical Psychology,* ed. D. Luce (John Wiley and Sons).

Tabesh, M., Mirström, M., Böhme, R. A., Lasota, M., Javaherian, Y., Agbotsoka-Guiter, T., et al. (2025). Question-based computational language approach outperform ratings scale in discriminating between anxiety and depression. *J. Anxiety Disord.* 112:103020. doi: 10.1016/j.janxdis.2025.103020

Tal, E. (2017). Calibration: modelling the measurement process. *Stud. History Philos. Sci.* 65, 33–45. doi: 10.1016/j.shpsa.2017.09.001

Tal, E. (2020). "Measurement in science," in *The Stanford Encyclopedia of Philosophy,* ed. E. N. Zalta (Metaphysics Research Lab, Stanford University). Available online at: https://plato.stanford.edu/archives/fall2020/entries/measurement-science (accessed October 21, 2023).

Teigen, K. H. (2018). The unbearable lightness of finger movements: commentary to Doliński. *Soc. Psychol. Bull.* 13, 1–6. doi: 10.5964/spb.v13i2.26110

Thomas, M. A. (2019). Mathematization, not neasurement: a critique of Stevens' scales of measurement. *J. Methods Measur Soc. Sci.* 10, 76–94. doi: 10.2458/v10i2.23785

Thurstone, L. L. (1928). Attitudes can be measured. *Am. J. Sociol.* 33, 529–554. doi: 10.1086/214483

Toomela, A. (2008). Variables in psychology: a critique of quantitative psychology. *Integr. Psychol. Behav. Sci.* 42, 245–265. doi: 10.1007/s12124-008-9059-6

Toomela, A. (2010). "Modern mainstream psychology is the best? Noncumulative, historically blind, fragmented, atheoretical," in *Methodological thinking in psychology: 60 years gone astray?* eds. A. Toomela and J. Valsiner (Charlotte: Information Age Publishers), 1–26.

Toomela, A., and Valsiner, J. (2010). *Methodological Thinking in Psychology : 60 Years Gone Astray?* Charlotte, NC: Information Age Publishing.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York, NY: Wiley.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511819322

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory Psychol.* 19, 579–599. doi: 10.1177/0959354309341926

Truijens, F. L. (2017). Do the numbers speak for themselves? A critical analysis of procedural objectivity in psychotherapeutic efficacy research. *Synthese* 194, 4721–4740. doi: 10.1007/s11229-016-1188-8

Truijens, F. L., Cornelis, S., Desmet, M., De Smet, M. M., and Meganck, R. (2019a). Validity beyond measurement: why psychometric validity is Insufficient for valid psychotherapy research. *Front. Psychol.* 10:532. doi: 10.3389/fpsyg.2019.00532

Truijens, F. L., Kimberly, V. N., Melissa, M., Mattias, D., and Meganck, R. (2022). How questionnaires shape experienced symptoms. A qualitative case comparison study of questionnaire administration in psychotherapy research. *Qualit. Res. Psychol.* 19, 806–830. doi: 10.1080/14780887.2021.1886383

Truijens, F. L., Mattias, D., Eva, D. C., Horanka, U., Bram, D., and Meganck, R. (2019b). When quantitative measures become a qualitative storybook: a phenomenological case analysis of validity and performativity of questionnaire administration in psychotherapy research. *Qual. Res. Psychol.* 19, 244–287. doi: 10.1080/14780887.2019.1579287

Uher, J. (2013). Personality psychology: lexical approaches, assessment methods, and trait concepts reveal only half of the story-Why it is time for a paradigm shift. *Integr. Psychol. Behav. Sci.* 47, 1–55. doi: 10.1007/s12124-013-9230-6

Uher, J. (2015a). "Agency enabled by the psyche: explorations using the transdisciplinary philosophy-of-science paradigm for research on individuals," in *Constraints of agency: Explorations of theory in everyday life. Annals of Theoretical Psychology,* eds. C. W. Gruber, M. G. Clark, S. H. Klempe, and J. Valsiner (New York: Springer International Publishing), 177–228. doi: 10.1007/978-3-319-10130-9_13

Uher, J. (2015b). Conceiving "personality": psychologist's challenges and basic fundamentals of the transdisciplinary philosophy-of-science paradigm for research on individuals. *Integr. Psychol. Behav. Sci.* 49, 398–458. doi: 10.1007/s12124-014-9283-1

Uher, J. (2015c). Developing "personality" taxonomies: metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integr. Psychol. Behav. Sci.* 49, 531–589. doi: 10.1007/s12124-014-9280-4

Uher, J. (2015d). Interpreting "personality" taxonomies: why previous models cannot capture individual-specific experiencing, behaviour, functioning and development. Major taxonomic tasks still lay ahead. *Integr. Psychol. Behav. Sci.* 49, 600–655. doi: 10.1007/s12124-014-9281-3

Uher, J. (2016a). "Exploring the workings of the Psyche: Metatheoretical and methodological foundations," in *Psychology as the science of human being: The Yokohama Manifesto,* eds. J. Valsiner, G. Marsico, N. Chaudhary, T. Sato, and V. Dazzani (New York: Springer International Publishing), 299–324. doi: 10.1007/978-3-319-21094-0_18

Uher, J. (2016b). What is behaviour? And (when) is language behaviour? A metatheoretical definition. *J. Theory Soc. Behav.* 46, 475–501. doi: 10.1111/jtsb.12104

Uher, J. (2018a). Quantitative data from rating scales: an epistemological and methodological enquiry. *Front. Psychol.* 9:2599. doi: 10.3389/fpsyg.2018.02599

Uher, J. (2018b). Taxonomic models of individual differences: a guide to transdisciplinary approaches. *Philos. Trans. R. Soc. B* 373:20170171. doi: 10.1098/rstb.2017.0171

Uher, J. (2019). Data generation methods across the empirical sciences: differences in the study phenomena's accessibility and the processes of data encoding. *Qual. Quan. Int. J. Methodol.* 53, 221–246. doi: 10.1007/s11135-018-0744-3

Uher, J. (2020a). Human uniqueness explored from the uniquely human perspective: epistemological and methodological challenges. *J. Theory Soc. Behav.* 50, 20–24. doi: 10.1111/jtsb.12232

Uher, J. (2020b). Measurement in metrology, psychology and social sciences: data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Qual. Quan. Int. J. Methodol.* 54, 975–1004. doi: 10.1007/s11135-020-00970-2

Uher, J. (2021a). Problematic research practices in psychology: misconceptions about data collection entail serious fallacies in data analysis. *Theory Psychol.* 31, 411–416. doi: 10.1177/09593543211014963

Uher, J. (2021b). Psychology's status as a science: peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integr. Psychol. Behav. Sci.* 55, 212–224. doi: 10.1007/s12124-020-09545-0

Uher, J. (2021c). Psychometrics is not measurement: unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.* 41, 58–84. doi: 10.1037/teo0000176

Uher, J. (2021d). Quantitative psychology under scrutiny: measurement requires not result-dependent but traceable data generation. *Pers. Individ. Dif.* 170:110205. doi: 10.1016/j.paid.2020.110205

Uher, J. (2022a). Functions of units, scales and quantitative data: Fundamental differences in numerical traceability between sciences. *Qual. Quan. Int. J. Methodol.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: a rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

Uher, J. (2023a). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Soc. Person. Psychol. Compass* 17:e12740. doi: 10.1111/spc3.12740

Uher, J. (2023b). What are constructs? Ontological nature, epistemological challenges, theoretical foundations and key sources of misunderstandings and confusions. *Psychol. Inquiry* 34, 280–290. doi: 10.1080/1047840X.2023.2274384

Uher, J. (2024). "Transdisciplinarity, complexity thinking and dialectics," in *The Routledge International Handbook of Dialectical Thinking,* eds. N. Shannon, M. Mascolo, and A. Belolutskaya (London: Routledge), 259–277. doi: 10.4324/9781003317340-21

Uher, J., Addessi, E., and Visalberghi, E. (2013a). Contextualised behavioural measurements of personality differences obtained in behavioural tests and social observations in adult capuchin monkeys (*Cebus apella*). *J. Res. Pers.* 47, 427–444. doi: 10.1016/j.jrp.2013.01.013

Uher, J., Arnulf, J. K., Barrett, P. T., Heene, M., Heine, J.-H., Martin, J., et al. (2025). Psychology's questionable research fundamentals (QRFs): key problems in quantitative psychology and psychological measurement beyond questionable research practices (QRPs). *Front. Psychol.* 16:1553028. doi: 10.3389/fpsyg.2025.1553028

Uher, J., and Visalberghi, E. (2016). Observations versus assessments of personality: a five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments. *J. Res. Pers.* 61, 61–79. doi: 10.1016/j.jrp.2016.02.003

Uher, J., Werner, C. S., and Gosselt, K. (2013b). From observations of individual behaviour to social representations of personality: developmental pathways, attribution biases, and limitations of questionnaire methods. *J. Res. Pers.* 47, 647–667. doi: 10.1016/j.jrp.2013.03.006

Valsiner, J. (1998). *The Guided Mind : A Sociogenetic Approach to Personality*. Cambridge, MA: Harvard University Press.

Valsiner, J. (2000). *Culture and Human Development*. London: Sage. doi: 10.4135/9781446217924

Valsiner, J. (2007). *Culture in Minds and* Societies: *Foundations of Cultural Psychology*. New York: Sage. doi: 10.4135/9788132108504

Valsiner, J. (2014a). *An invitation to Cultural Psychology*. New York: SAGE Publications. doi: 10.4135/9781473905986

Valsiner, J. (2014b). Needed for cultural psychology: methodology in a new key. *Cult. Psychol.* 20, 3–30. doi: 10.1177/1354067X13515941

Valsiner, J. (2017). *From Methodology to Methods in Human Psychology*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-61064-1

Valsiner, J., Diriwächter, R., and Sauck, C. (2005). "Diversity in unity: standard questions and nonstandard interpretations," in *Science and medicine in dialogue,* eds. R. Bibace, J. Laird, K. Noller, and J. Valsiner (Westport, CT: Praeger-Greenwood), 289–307. doi: 10.5040/9798216011491.ch-018

van der Maas, H., Kan, K.-J., and Borsboom, D. (2014). Intelligence is what the intelligence test measures. *Seriously. J. Intell.* 2, 12–15. doi: 10.3390/jintelligence2010012

van Fraassen, B. C. (2008). *Scientific* Representation: *Paradoxes of Perspective*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199278220.001.0001

van Geert, P. (2011). The contribution of complex dynamic systems to development. *Child Dev. Perspect.* 5, 273–278. doi: 10.1111/j.1750-8606.2011.00197.x

Vanier, A., Oort, F. J., McClimans, L., Ow, N., Gulek, B. G., Böhnke, J. R., et al. (2021). Response shift in patient-reported outcomes: definition, theory, and a revised model. *Qual. Life Res.* 30, 3309–3322. doi: 10.1007/s11136-021-02846-w

Velleman, P. F., and Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *Am. Stat.* 47, 65–72. doi: 10.1080/00031305.1993.10475938

Vessonen, E. (2017). Psychometrics versus representational theory of measurement. *Philos. Soc. Sci.* 47, 330–350. doi: 10.1177/0048393117705299

von Eye, A., and Bogat, G. A. (2006). Person-oriented and variable-oriented research: concepts, results, and development. *Merrill Palmer Q.* 52, 390–420. doi: 10.1353/mpq.2006.0032

von Neumann, J. (1955). *Mathematical foundations of quantum mechanics [originally published as Mathematische Grundlagen der Quantenmechanik in 1935]*. Princeton, NJ: Princeton University Press.

Vygotsky, L. S. (1962). *Thought and Language*. Cambridge, MA: MIT Press. doi: 10.1037/11193-000

Wagoner, B., and Valsiner, J. (2005). "Rating tasks in psychology: From a static ontology to a dialogical synthesis of meaning," in *Contemporary Theorizing in Psychology: Global Perspectives,* eds. A. Gülerce, I. Hofmeister, G. Saunders, and J. Kaye (Toronto, Canada: Captus), 197–213.

Watts, A. W., and Watts, M. (1996). *Myth and religion. The edited transcripts*. Tuttle Publishing. Available online at: https://www.organism.earth/library/document/not-what-should-be-but-what-is (accessed November 02, 2024).

Weber, M. (1949). *On the Methodology of the Social Sciences*. New York: Free Press.

Westen, D. (1996). A model and a method for uncovering the nomothetic from the idiographic: an alternative to the Five-Factor Model. *J. Res. Pers.* 30, 400–413. doi: 10.1006/jrpe.1996.0028

Wojciszke, B., and Bocian, K. (2018). Bad methods drive out good: the curse of imagination in social psychology research. *Soc. Psychol. Bull.* 13, 1–6. doi: 10.5964/spb.v13i2.26062

Wulff, D. U., and Mata, R. (2023). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nat. Hum. Behav.* 9, 944–954. doi: 10.1038/s41562-024-02089-y

Wundt, W. (1896). *Grundriss der Psychologie [Outlines of Psychology]*. Stuttgart: Körner.

Yildiz, T. (2025). The minds we make: a philosophical inquiry into theory of mind and artificial intelligence. *Integr. Psychol. Behav. Sci.* 59:10. doi: 10.1007/s12124-024-09876-2

frontiers | Frontiers in Psychology

TYPE Review
PUBLISHED 25 August 2025
DOI 10.3389/fpsyg.2025.1553028



OPEN ACCESS

EDITED BY
Fernando Marmolejo-Ramos,
Flinders University, Australia

REVIEWED BY
David Trafimow,
New Mexico State University, United States
James W. Grice,
Oklahoma State University, United States

*CORRESPONDENCE
Jan Ketil Arnulf
jan.k.arnulf@bi.no

†Deceased

RECEIVED 29 December 2024
ACCEPTED 01 July 2025
PUBLISHED 25 August 2025

CITATION
Uher J, Arnulf JK, Barrett PT, Heene M, Heine J-H, Martin J, Mazur LB, McGann M, Mislevy RJ, Speelman C, Toomela A and Weber R (2025) Psychology's Questionable Research Fundamentals (QRFs): Key problems in quantitative psychology and psychological measurement beyond Questionable Research Practices (QRPs). *Front. Psychol.* 16:1553028. doi: 10.3389/fpsyg.2025.1553028



# Psychology's Questionable Research Fundamentals (QRFs): Key problems in quantitative psychology and psychological measurement beyond Questionable Research Practices (QRPs)

Jana Uher [1], Jan Ketil Arnulf [2,3]*, Paul T. Barrett [4†], Moritz Heene [5], Jörg-Henrik Heine [6], Jack Martin [7], Lucas B. Mazur [8,9], Marek McGann [10], Robert J. Mislevy [11†], Craig Speelman [12], Aaro Toomela [13] and Ron Weber [14,15]

[1]School of Human Sciences, University of Greenwich, London, United Kingdom, [2]BI Norwegian Business School, Oslo, Norway, [3]Norwegian Defence University College, Oslo, Norway, [4]Advanced Projects R&D Ltd, Auckland, New Zealand, [5]Ludwig-Maximilians-Universität München, Munich, Germany, [6]University of the Bundeswehr Munich, Neubiberg, Germany, [7]Department of Psychology, Simon Fraser University, Burnaby, BC, Canada, [8]Jagiellonian University, Kraków, Poland, [9]Sigmund Freud University Berlin, Berlin, Germany, [10]Mary Immaculate College, Limerick, Ireland, [11]University of Maryland, Baltimore, MD, United States, [12]Edith Cowan University, Joondalup, WA, Australia, [13]Tallinn University, Tallinn, Estonia, [14]Faculty of Information Technology, Monash University, Melbourne, VIC, Australia, [15]School of Business, The University of Queensland, Brisbane, QLD, Australia

Psychology's crises (e.g., replicability, generalisability) are currently believed to derive from Questionable Research Practices (QRPs), thus scientific misconduct. Just improving the same practices, however, cannot tackle the root causes of psychology's problems—the Questionable Research Fundamentals (QRFs) of many of its theories, concepts, approaches and methods (e.g., psychometrics), which are grounded in their insufficiently elaborated underlying philosophies of science. Key problems of psychological measurement are critically explored from independent perspectives involving various fields of expertise and lines of research that are well established but still hardly known in mainstream psychology. This comprehensive multi-perspectival review presents diverse philosophies of science that are used in quantitative psychology and pinpoints four major areas of development. (1) Psychology must advance its general philosophy of science (esp. ontology, epistemology, methodology) and elaborate coherent paradigms. (2) Quantitative psychologists must elaborate the philosophy-of-science fundamentals of specific theories, approaches and methods that are appropriate for enabling quantitative research and for implementing genuine analogues of measurement in psychology, considering its study phenomena's peculiarities (e.g., higher-order complexity, non-ergodicity). (3) Psychologists must heed the epistemic necessity to logically distinguish between the study phenomena (e.g., participants' beliefs) and the means used for their exploration (e.g., descriptions of beliefs in items) to avoid confusing ontological with epistemological concepts—psychologists' cardinal error. This requires an increased awareness of the complexities of human language (e.g.,

inbuilt semantics) and of the intricacies that these entail for scientific inquiry. (4) Epistemically justified strategies for generalising findings across unique individuals must be established using case-by-case based (not sample-based) nomothetic approaches, implemented through individual-/person-oriented (not variable-oriented) analyses. This is crucial to avoid the mathematical-statistical errors that are inherent to quantitative psychologists' common sample-to-individual inferences (e.g., ergodic fallacy) as well as to enable causal analyses of possibly underlying structures and processes. Concluding, just minimising scientific misconduct, as currently believed, and exploiting language-based algorithms (NLP, LLMs) without considering the intricacies of human language will only perpetuate psychology's crises. Rethinking psychology as a science and advancing its philosophy-of-science theories as necessary fundamentals to integrate its fragmented empirical database and lines of research requires open, honest and self-critical debates that prioritise scientific integrity over expediency.

KEYWORDS

**measurement, quantitative psychology, psychometrics, language models, ontology, epistemology, methodology, semantics**

## Questionable Research Practices (QRPs): Surface-level symptoms obscuring fundamental problems still largely overlooked

Psychology's crises in replicability, validity and generalisability reflect a lack of scientific and societal confidence in its research findings (Newton and Baird, 2016; Open Science Collaboration, 2015; Schimmack, 2021; Yarkoni, 2022). Many psychologists attribute these crises to the improper application of established research methods—termed *Questionable Research Practices* (QRPs; John et al., 2012). These involve hypothesising after the results are known (HARKing), analysing data relentlessly to obtain statistically significant results that support the researchers' hypotheses (*p*-hacking), testing statistical associations of randomly combined variables without any theoretical hypotheses (fishing) and other questionable practices (Andrade, 2021; Earp and Trafimow, 2015). For the meticulous method expert, these flaws are readily identifiable, as are their remedies—larger samples, more robust statistics, more data transparency (open science, preregistration; e.g., Nosek et al., 2015; Zwaan et al., 2017). Thus, do psychology's crises arise just because psychologists are more prone to scientific misconduct than scholars in other disciplines?

## Psychology's Questionable Research Fundamentals (QRFs)

Most quantitative psychologists use approaches (e.g., research designs) and methods of empirical inquiry (e.g., rating 'scales', statistical analyses) that are well-established in the field. Its leaders focus on advancing and applying these standards meticulously, wary of Questionable Research Practices (QRPs). We believe, however, that psychology's recurring crises cannot be overcome by just improving the same practices. We believe a fundamental rethinking is necessary.

Like all scientific activities, the approaches and methods of quantitative psychology are built on *presumptions*, which inform their rationales and operations—thus, on ideas that are taken for granted with confident belief until it can be proved otherwise. All theories, approaches and methods are also built on beliefs about what exists for us to know about, how we can generate knowledge and what is possible for us to know and in what ways. These *presuppositions*—fundamental, often unstated beliefs that underlie a system of knowledge—guide the decisions that any empirical scientist must make about what to study, what to regard as fact, what questions to address, what procedures and operations to use for exploring these as well as how to interpret results (Collingwood, 1940; Fleck, 1935/1979; Kuhn, 1962/1970; Uher, 2013; Valsiner, 2012; Weber, 1949). These fundamental beliefs may not be considered explicitly by everyone doing quantitative research. Still, as generalised views on how to do science, they influence all scientific activities in a field.

We have come together as scholars from different backgrounds and disciplines to critically reflect on quantitative psychology's research fundamentals and its current problems because a classical review, which always provides just a few authors' views, is insufficient. There are also no criteria on which a classical review could be based—because what is required is a rethinking of the very fundamentals on which many established practices are built. We therefore do not discuss ways to improve specific quantitative methods and approaches (e.g., statistical modelling) or their meticulous application, as commonly done. In our view, questionable research practices are just surface-level symptoms that distract from and obscure the root causes of psychology's crises—the *Questionable Research Fundamentals (QRFs)* of many of its theories, concepts, approaches and methods. Therefore, our focus is on making explicit and scrutinising the fundamental principles and rationales on which quantitative psychology is currently built. We outline alternative ones on which it could and should be built in the future.

This also requires critically analysing and elaborating the underlying *philosophies and theories of science*. Their relevance for quantitative psychology, however, is often overlooked. Many regard them as a mere specialist field, studied by just a small minority of psychologists. But all scientific research is based on a philosophy and theory of science—otherwise it would not be science. Specifically, all science is aimed at understanding the 'world'—that is, it has a basic ontological orientation. All science is also concerned with our knowledge of this 'world'—thus, it also has a basic epistemological orientation. Now, what does this involve?

## Philosophy and theory of science—The fundamentals of scientific inquiry: Ontology, epistemology and methodology

Philosophy of science is concerned with the most fundamental questions of scientific inquiry. It involves ontology, epistemology, methodology and further branches of philosophy. *Ontology*, the philosophy and theory of being, is concerned with the most fundamental kinds of being that may be taken to exist, especially with their categorisation, structures and relations. *Epistemology*[1], the philosophy and theory of knowing, is concerned with the nature and scope of knowledge that we can generate about specific kinds of being. This involves, amongst others, the justification of knowledge claims, concepts of 'truth', logic and rationality. Epistemological presuppositions influence how researchers frame and design their research as well as how they view the relation between themselves and their objects of research—between the researcher and the researched. *Methodology*, the philosophy and theory of methods, in turn, connects abstract ontology and epistemology with empirical research. It provides justification for why specific procedures and operations (methods), but not others, are suited to explore specific objects of research and specific questions (Ali, 2023; Hartmann, 1964; Mertens, 2023; Poli and Seibt, 2010; Uher, 2022b, 2025; Valsiner, 2017).

In psychology and other sciences, many different ontologies, epistemologies and methodologies have been developed for different objects of research, different aims and purposes, and from different worldviews. This leads to pronounced differences in the specific ways of doing science that are pursued in a field—thus, to different paradigms.

## Research paradigms in psychology: Diversity in the ways of doing science

A *paradigm* is a distinct framework that provides a coherent set of theories, models, concepts, terms, instruments and practices that are often considered conventional in a field and that build on a specific worldview and specific presuppositions and values. Paradigms may arise in a field from a single scholar's research that serves as an exemplar for solving fundamental problems (e.g., Newton's). Its successes promote consensus among other scholars and agreement on the framework on which it is based. Often, however, paradigms emerge gradually over time from theoretical, methodical and empirical advances that are made by many scholars in a field, each exploring specific problems and questions. Some paradigms are already more elaborated and coherent in their philosophical fundamentals, whereas others are more implicit and still awaiting coherent elaboration. Paradigm-specific jargon, however, often makes it difficult to immediately see commonalities and differences between paradigms. Their elaboration, however, is important to recognise the implications that paradigmatic differences have for empirical research. This is also necessary to understand the different quality criteria and standards of evaluation that apply to different paradigms and that often preclude direct comparisons (incommensurability; Bird, 2022; Kuhn, 1962/1970).

These complex fundamentals are worth exploring in their own rights (Ali, 2023; Fahrenberg, 2013, 2015; Holzkamp, 1983; Jovanović, 2022; Mertens, 2023; Toomela and Valisiner, 2010; Uher, 2018a, 2021c; Valsiner, 2017). But we do not aim to systematically elaborate them here and such is not necessary for our analyses. Like all scholars, we have our specialisations. Not all of us are scrutinising and elaborating the philosophy-of-science fundamentals of theories, concepts, terms, approaches and methods. Still, in this article, we want to create and increase awareness of the philosophical and theoretical dimensions underlying quantitative research in psychology and the disciplinary crises that it encounters. Therefore, we highlight important points to enable a more in-depth understanding of the current problems and their underlying Questionable Research Fundamentals (QRFs).

For this purpose, we aim to provide a more comprehensive overview of independent perspectives that can and should be taken on quantitative psychology's current status and development as a science. These involve many established lines of research from smaller communities of research and practice, often published outside of mainstream journals and thus, outside most psychologists' focus. As experts in our respective fields, we independently provide a critical reflection of what we see as quantitative psychology's main problems and what as the key tasks that must be tackled. We present solutions that have already been developed, explain their fundamentals and direct readers to key publications. This highlights another crucial point.

## Diverse perspectives, philosophies and theories of science required in psychology

In any given discipline, there can be no single one-and-only right way of doing science—especially not in psychology, given that it explores phenomena as diverse as brain morphology, physiology, behaviour, experience, social interaction, language and other socio-cultural products of the human mind (Uher, 2021c). This highlights a further key point. Diverse perspectives, philosophies and theories of science are not just possible in psychology—they are even necessary, also in quantitative psychology. This requires, first and foremost, awareness and efforts to make basic presuppositions

1 Ontology and epistemology as well as their relation are variously defined. Still, the two should not be confounded. Assuming that both are interdependent, ontology can state about epistemology that concepts, theories, presumptions and beliefs are (scholars') psychical (e.g., mental) phenomena by their ontological nature. Hence, knowledge of a being is a state of being itself. Epistemology, in turn, can say about ontology that knowledge of the structure of beings is a kind of knowledge itself (Poli, 2001; Uher, 2023b).

explicit and thus, accessible to elaboration and analysis. This also requires scholars to be tolerant and open to different perspectives to be able to not just pinpoint and critically discuss differences but also to identify communalities—because these may not always be obvious (Uher, 2024).

Indeed, although each of our independent contributions has its own focus and rationale, they also show systematic connections with one another, thereby creating a poly-perspectival and more comprehensive overview than any review by single authors could provide. With our compilation of different perspectives, ways of thinking and doing science, we also aim to foster the scientific spirit of an open debate in which we can make explicit our most basic philosophical presuppositions, challenge established concepts, theories and practices, advance novel ways of thinking and exchange controversially—yet constructively and collegially—about scientific psychology.

## Outline of this article

Our critical analyses are grouped into four main areas that cover different topics, problems and research questions and that, in our view, require remediation, elaboration and further development (Figure 1). Topic 1 starts by exploring quantitative psychology as a science. *Lucas Mazur* reflects on psychology's struggle with its scientific status and on the problems, promises and perils of scientism. *Aaro Toomela* elaborates on what science actually is as well as on the imperative to advance psychology's ontology, epistemology and methodology and to align them to one another to develop coherent paradigms. *Jack Martin* reminds us of the inherent contextuality of human experience that makes up personhood and draws conclusions for quantitative and experimental psychology.

Topic 2 is devoted to the specific epistemological, methodological and theoretical foundations of psychometrics and psychological 'measurement', highlighting fundamental differences to physical measurement that are still not well considered. *Jana Uher* explores the conceptual problems entailed by psychology's operationalist definition of 'measurement' and quantitative data generation with rating 'scales', and highlights incompatibilities in the epistemological framework on which psychometrics is built. *Jörg-Henrik Heine* and *Moritz Heene* locate the failed promises of psychological 'measurement' in the impossibility to establish one–to–one relations between the phenomenological object domain and the mathematical metric space of positive real numbers. *Paul Barrett* concurs that, without meeting the axioms of quantity and the human mind's peculiarities, quantitative psychology cannot implement genuine measurement processes. He highlights that the increasingly popular use of generative language algorithms cannot solve these fundamental problems. *Robert Mislevy* derives from the contextuality of human experience and learning a socio-cognitive approach that re-conceptualises the theoretical and philosophical framework that is necessary for making justified inferences from quantitative educational assessments in applied settings, while avoiding conceptual errors inherent in current conceptions. *Jana Uher* demonstrates that statistics and measurement are different scientific activities designed for different epistemic purposes. She specifies basic criteria and methodological principles and explains the system of modelling relations that are epistemically necessary for establishing genuine analogues of measurement in psychology.

Topic 3 explores the intricate relations between psychologists' study phenomena (e.g., participants' beliefs) and their means for investigating these phenomena (e.g., descriptions of beliefs in rating 'scales' and models). *Jana Uher* highlights that their logical distinction (in each study) is an epistemic necessity to avoid conflating ontological with epistemological concepts—psychologists' cardinal error. *Jan Ketil Arnulf* therefore demands a more critical reflection on the role of human language in scientific inquiry. He demonstrates the epistemic necessity to distinguish between empirical and semantic research problems by showing that the inbuilt semantics of item statements, analysed through natural language algorithms, produces results similar to those obtained from empirical rating studies. *Ron Weber* analyses the ontology of construct–indicator and indicator–instrument relationships and introduces novel ontological concepts to analyse the applicability of constructs and their operationalisations (indicators) to different subsamples of populations, highlighting their implications for instrument development.

Topic 4 critically analyses psychology's approaches for generalising findings across unique individuals. It demonstrates that psychology's default use of sample-level statistics to explore individual-level phenomena ignores the mathematical-statistical foundations of such inferences (ergodic theory), the non-ergodicity of psychology's study phenomena as well as the peculiarities of complex living systems and therefore entails various inferential fallacies. *Craig Speelman* and *Marek McGann* highlight that the common sample-to-individual inferences build on the ergodic fallacy, thereby contributing to psychology's inferential and reproducibility problems, and they present pervasiveness analysis as an alternative approach. *Jana Uher* shows that, to avoid fallacies when making sample-to-individual inferences, psychology must advance case-by-case based (not group-based) nomothetic approaches, implemented through individual-/person-oriented (not variable-oriented) analyses. This is essential for identifying actual commonalities and differences among individuals as well as for enabling causal analyses to unravel (possibly) underlying structures and processes.

We close with general conclusions and future directions, highlighting that just minimising scientific misconduct and exploiting the new generative language algorithms to design 'scales' and constructs, as increasingly done, will not remedy but only intensify psychology's problems and crises. Instead, tackling psychology's Questionable Research Fundamentals (QRFs) requires critical self-reflection and a fundamental rethinking of doing science in psychology.

To give new impetus to the current debates, we now discuss each of the four areas of development that we have identified (Topics 1 to 4) and present various independent perspectives, each focussed on specific problems and research questions. We analyse commonalities and differences of established and alternative ways of doing science in quantitative psychology, highlight their underlying philosophies of science, pinpoint key issues and provide novel insights.


FIGURE 1
Psychology's Questionable Research Fundamentals (QRFs): Four main areas of development.

## Topic 1: Quantitative psychology as a science: Key assumptions and the necessary philosophy-of-science fundamentals

Quantitative psychology has been developed in response to continued doubts, first voiced by Immanuel Kant in the 18th century, on psychology's ability to become an exact experimental—thus, a 'real'—science. But even in the 21st century, psychology is still struggling with its status as a scientific discipline (Uher, 2021c).

### Why does psychology continue to struggle with its scientific status? The blinding promises and perils of scientism

In the attempt to make the field indubitably "scientific", quantitative psychologists often end up embracing *scientism*, the belief that "only scientific knowledge counts as real knowledge" (Williams, 2015, p. 6). Why do many continue to believe in the promises of scientism, while ignoring the problems and even perils that it brings? In his line of research on psychology's history and philosophy of science, Lucas Mazur explored this question conceptually (Mazur, 2015, 2017, 2021, 2024a; Mazur and Watzlawik, 2016). In his empirical research, he encourages interpretive, anti-naturalistic (treating psychological phenomena not like natural facts), dynamic and contextualised approaches—even when making use of quantitative methods (Mazur et al., 2022; Mazur and Sticksel, 2021; Mazur, 2022, 2024b,c).

#### The problems

In psychology, there is a persistent blindness to the problems of scientism. These include *quantificationism* (viewing quantitative information as generally superior to qualitative information), *naturalism* (viewing research data as raw, objective 'natural' facts in need of little or no interpretation), *statisticism* (viewing statistics as a complete or sufficient basis for scientific methodology) and *psychologism* (reducing thought and knowledge to internal psychological characteristics of individual minds), amongst others (Lamiell, 2018, 2019a; Sugarman, 2017; Uher, 2022b). From time

to time, the problems become so undeniable as to demand a response (e.g., the replication crisis). But, at those moments, many quantitative psychologists perennially respond with *more of the same* (e.g., open science, robust statistics)—in effect, "kicking the can down the road" (Steinmetz, 2005; Tolman, 1992). As Valsiner and Brinkmann (2016) suggested,

> "[it] cannot be the case that this unfortunate situation occurs only due to the intellectual transformations within the history of psychology itself. There must be some societal catalytic process for the meta-theoretical blindness in the field" (p. 87).

For the fact that many researchers do not look the problems of scientism squarely in the face, there is, amongst many others (Uher, 2022b), a two-sided societal reason. This is the blinding power of scientism, particularly the belief that quantification is a step towards prediction and ultimately control (Hacking, 1990; Porter, 1995). On the one side, many researchers are so thoroughly pulled towards scientism that they do not reflect on this choice of direction. Indeed, they do not even see it as a choice but as the only way to go—rendering the matter "too obvious" to warrant consideration and its potential loss as their lodestar too disorienting. On the other side, if they paused for serious reflection, they would see a vision that forces them to close their eyes in disgust, or even horror, which is likewise deterring proper reflection. This tension creates a form of collective avoidance that perpetuates problematic meta-theoretical and methodological assumptions (Mazur, 2021).

## The promises

The gravitational forces of quantification and mechanistic causality have become so powerful that they distort many researchers' very perception of 'reality', as reflected in the belief that "science is the only path to understanding" (Gnatt, 2018). This view has become deeply entrenched in quantitative psychology, where it is widely believed that human experience and behaviour can be reduced to measurable, predictable units (Michell, 2022). The contemporary emphasis on optimisation—both in academia and society at large—exemplifies this mindset. It is often reinforced, paradoxically, even by attempts to resist this trend: calls to "unplug" or "slow down" frequently come packaged in the language of optimisation as quantifiable steps leading to quantifiable benefits.

Over a century ago, social theorists presciently identified that this shift towards quantification was part of the broader process whereby scientists become tools of their own tools (Danziger, 1990; Daston and Galison, 2007; Poovey, 1998; Valsiner, 2007, 2012). Max Weber (1904–05/1992) noted how commitment to non-calculable goals was increasingly viewed as irrational. Durkheim (1893/1984) recognised the cultural dominance of this new rationality but struggled to envision alternatives. Simmel (1900/1978, p. 443) similarly observed "the growing preponderance of the category of quantity over that of quality, or more precisely the tendency to dissolve quality into quantity". Today, this tendency has become further intensified under the auspices of neoliberal consumerism, which further privileges quantity over quality (Sugarman, 2017). Therefore, many researchers keep their eyes fixed on the horizon of 'progress' that this quantity-focused worldview promises, driven simultaneously by an unspoken anxiety that any deviation from it might halt humanity's collective march forward.

## The perils

The disturbing implications of scientism become apparent when we examine its logical conclusions. As Maslow (1966, p. 75) noted, in scientism, "the blueprints are more real than the houses. The maps are more real than the territory". For many psychologists, this is already a disturbing denial of our humanity. However, the prioritisation of measurement, prediction and control at the cost of all that does not fit the mould points in even darker directions. This became apparent, for example, in *classical positivism*, which is built on the presupposition that the social 'world' can be explored just like the natural 'world' through observation, experimentation and measurement by independent researchers who work objectively and separated from their own values. The founders of classical positivism, Henri de Saint Simon and Auguste Comte, even called people resistant to positivism "parasites" and mere "dung-producers", arguing that they "transmit to their successors no equivalent for what they received from their predecessors" and therefore "should be treated like cattle" (de Lubac, 1995).

This dehumanising language and logic haunt the boundaries of both science and morality—of what *could* and what *should* be done. It is no coincidence that such thinking appears in dystopian works like Animal Farm, Brave New World as well as Frankenstein ("the modern Prometheus"). It has also found its expression in eugenics, communism and National Socialism. Meanwhile, embracing scientific approaches to understanding human nature has become second nature to many researchers. However, when this embrace becomes exclusive and dismissive of other perspectives, researchers risk creating the very scenarios that science fiction—and history—have long warned against. The tendency to quickly pass over figures like Comte and Saint-Simon in psychological teaching and textbooks may perhaps reflect not just the *naturalisation* of social science—its treatment like a natural science—but also an unconscious recoil from its more troubling implications.

## The prospects

To be swept away by a scientistic vision of humanity is to soar on the wings of Icarus. Once in the air, many either keep their eyes focused on the blinding sun towards which they are heading (the promises), or they keep them closed in terror before the fall (the perils). Either way, they do not want to *see*. Below one can hear the flapping of the perilously glued-on wings:

> "If I could only discover some external indicator of, for example, happiness or anxiety, some litmus paper test of the subjective, I would be a very happy man. But happiness and anxiety now exist in the absence of such objective tests. It is the denial of this existence that I consider so silly that I won't bother arguing about it. Anyone who tells me that my emotions and my desires don't exist is in effect, telling me that I don't exist" (Maslow, 1966, p. 47).

Beneath the dismissal of this "silly" suggestion, one can hear unease, even dread, but also desire—the simultaneous allure of, and

repulsion at, scientistic thinking. By contrast, social psychologist Gustaw Ichheiser wrote:

> "[S]ocial scientists should, in my opinion, not aspire to be as 'scientific' and 'exact' as physicists or mathematicians, but should cheerfully accept the fact that what they are doing belongs to the twilight zone between science and literature." (Ichheiser in a Letter in 1967, cited in Rudmin et al., 1987, p. 171).

This perspective suggests a practical path forward: integrating insights from the humanities more explicitly and thoroughly into psychological inquiry (Aeschliman, 1998; Bruner, 1990; Freeman, 2024; Mazur, 2015, 2017, 2021, 2024a; Sugarman and Martin, 2020). This is not a rejection of science. It is a recognition that—even after scientific methods have been applied in their proper scope to a limited range of phenomena in psychology (Mazur and Watzlawik, 2016; Taylor, 1985)—there remains much to study from other points of view and via other methods of investigation. While echoing the warnings against unreflective quantification, including against the faulty assumption that psychometrics could enable genuine measurement (Uher, 2021a), this is not a rejection of the thoughtful use of numbers as meaningful depictions of psychological phenomena (Mazur, 2022, 2024b). Indeed, both quantitative and qualitative methods can be both useful and problematic (Bevir and Blakely, 2018; Holzkamp, 2013). Psychologists do not even have to stop trying to positively impact the social 'world' around—after all, most of what is thought of as "psychological" already involves active engagement with that social 'world' (Ichheiser, 1943; Smedslund, 2016; Wittgenstein, 1953). This, however, is a reminder of how the temptations of power and control—which in psychology take the form of scientism—can blind many psychologists to the perpetual challenge of human hubris.

> "Let me warn you, Icarus, to take the middle way, in case the moisture weighs down your wings, if you fly too low, or if you go too high, the sun scorches them. Travel between the extremes." (Daedalus to his son Icarus)

The humanities, such as history, philosophy and literature but also rhetoric, music, the performing and visual arts, religious studies and theology, can help psychology to break free from the chains of scientism—from the desire for, and fear of, what researchers (mis)take to be scientific control. A more open-minded interweaving of fields will allow psychology to more richly understand, appreciate and wonder at the human condition.

This can and should entail systematic elaborations also of psychology's philosophy and theory of science. More and more psychologists are exploring epistemological and methodological issues as well as ontological questions, each with their specific focus on specific research questions and from their specific perspectives. At some point, the different elements of scientific inquiry used in a line of research should be elaborated and coherently aligned with one another and with the specific presumptions, beliefs and values on which they are based. This means that the specific epistemological approach used in a line of research should correspond to the specific ontological presumptions made and both should inform the corresponding methodology to guide the development of suitable methods (Al-Ababneh, 2020; Ali, 2023; Mertens, 2023). An example of such a coherently elaborated philosophy of science is the structural-systemic paradigm. This paradigm also opens a more fundamental perspective on psychology's crises, which goes much beyond the currently discussed surface-level symptoms of problems in replicability, validity and generalisability.

## The crisis still overlooked: Psychology's ontology, epistemology and methodology must be grounded in a structural-systemic paradigm

In his line of research on the ontological, epistemological and methodological foundations of psychology, Aaro Toomela highlighted that psychology's crisis is much more profound than currently considered. In fact, it is a *crisis in science*—defined as a situation where there is no generally accepted system of science (Vygotsky, 1982, p. 373, Vygotsky, 1997). Indeed, psychology is divided into mainstream psychology, which is pursued by the majority, and non-mainstream psychology, which challenges ontological, epistemological and methodological principles that are generally accepted by the mainstream (Toomela, 2014a, 2019).

Any science prospers best through collective efforts—through working as a global team. Scientific progress through collaboration is hindered, however, when it requires the discovery of novel questions that entail entirely novel perspectives on the object of research (Toomela, 2007b). This process is stretched over time. First, novel questions must be discovered and justified by individual scholars. When the questions are important, they must form groups of like-minded scholars who take the questions seriously and start developing new research approaches. Thereafter, it may still take considerable time before the importance of these novel questions and the novel approaches for answering them will be recognised by mainstream scholars.

Where is psychology now? There already is a set of novel questions about and novel perspectives on the general scientific worldview of mainstream psychology. These novel questions, as well as convincing approaches for tackling them, are increasingly discussed by various groups of non-mainstream scholars. But they are still largely ignored by mainstream psychologists. These questions concern the most basic principles of science—its ontology, epistemology and methodology. But first, what is science?

### What is science? And what is scientific understanding?

First, it is important to acknowledge that science is not necessary for achieving knowledge. Moreover, all knowledge about the 'world' is acquired only from information obtained directly through the sensory organs (in humans and animals alike). Most of the 'world', however, is not directly accessible with our human-specific senses. To understand the essence of science, it is necessary to distinguish between these two aspects of the material 'world'. Science came into being when humanity began to study those parts of the 'world' that are not accessible through our senses: science aims at understanding the 'world' that is not sensorily accessible in order to explain the 'world' that we can perceive with our senses (Toomela, 2022). Importantly, things and phenomena that appear

to be identical in our senses can sometimes be different in some aspects that we cannot sensorily perceive. Vice versa, things or phenomena that differ in our sensory perception may sometimes have common characteristics that are imperceptible to us. The essence of scientific methods is to help us discover such aspects of the 'world' that may causally underlie the directly perceivable and which may thus help us to explain the 'world' as it appears to us. Research methods that do not allow us to describe the parts of the 'world' that are inaccessible to our human senses therefore do not help us to advance our scientific knowledge.

## Scientific understanding of the human psyche requires a unifying ontological theory

Almost a century ago already, Vygotsky provided convincing arguments that psychology cannot become a true science without a *general-unifying theory* (Vygotsky, 1982; also Toomela, 2007c, 2014b, 2017)—an ontological theory of what the psyche is. The psyche as a whole can be defined as "a specifically organised form of living matter. Its purposeful behaviour in anticipating environmental changes that are harmful [or beneficial] for itself as a whole is based on individual experience" (Toomela, 2020, p. 29). This whole can be distinguished into parts at different levels of analysis. At the most general level, the psyche can be distinguished into the psychical individual and that part of the environment to which it relates (called the psychical environment; Toomela, 2020, also Koffka, 1935). In the psychical individual, further interrelated parts can be distinguished. Luria (1973) showed that the true material parts of the psyche are the different brain regions each with their unique function. *Vygotsky's theory of higher psychical functions* explains how the human psyche emerges when cultural signs become part of the structure of an individual's psychical system, which underlies its psychical processes (Toomela, 2016b; Vygotsky, 1994). Hence, within this general ontological theory, the psyche is defined as a structural system—as a whole. Such a theory is crucial to understand the essence of the human nature.

*Structural-systemic ontologies*—that is, presuppositions that the material 'world' is composed of hierarchies of interrelated parts that form qualitatively distinguishable whole structures at certain levels of analysis—are used in other sciences as well. Chemistry, for example, conceptualises molecules with different qualities, atoms as parts of molecules as well as the molecules' structure and their composition of atoms. Some molecules, called isomers, are composed of identical sets of atoms, but these are arranged in different relations from which qualitatively different molecules emerge as structured wholes. When we ontologically assume that the 'world' is systemically organised in interrelated structures in which parts are forming qualitatively different wholes, then an epistemology must be defined that corresponds to that ontology. Accordingly, the aim of science is to construct structural-systemic knowledge about that 'world'.

## Psychology requires a structural-systemic epistemology

Mainstream (quantitative) psychology pursues knowledge about generalised patterns in large data sets and (mostly) linear cause-effect relationships. But with these approaches, there is no way to discover the parts and processes of the psyche as well as the specific kinds of relationships between them and from which the particular properties of the psyche as a whole emerge (Toomela, 2020). What is required is a more powerful epistemology that Toomela called *structural-systemic* (Toomela, 2003, 2009a, 2012, 2014d, 2015, 2016a, 2019). This epistemology was pursued by several scholars in the history of psychology (e.g., Luria, 1973; Vygotsky, 1994; Werner, 1948; Wundt, 1897). Many further theories with various concepts of "system" and "structure" were developed in different sciences (see Ramage and Shipp, 2020). Hence, there is not just one but many structuralist or systems epistemologies. Therefore, it is necessary to define what specific theory is followed in a given line of research.

In Toomela's *structural-systemic epistemology*, science is aimed at constructing knowledge about the part–whole structures of the things or phenomena studied. In this approach, scientific understanding provides answers to three main questions: What is the studied whole? What are the parts of the whole? And in which relationships are these parts? The origin of this epistemology can be traced back to Aristotle who suggested that knowledge is about causal structures of the 'world'. He distinguished four *complementary kinds of causes*, nowadays called *material* (what are the parts), *formal* (what is the whole), *efficient* (what makes a change happen) and *final* (why does a change happen).

Today's mainstream psychology, by contrast, relies on a simplified Cartesian-Humean understanding of causality where only efficient causality is believed to be knowable. The Aristotelian perspective, however, shows that, to understand causality, all causes must be known. Specifically, the parts of a whole—its *material cause*—underlie what the whole is. Therefore, the whole cannot be understood without knowing the material cause because changes in the parts inevitably lead to the changes of the whole that is composed of and emerges from the parts. The whole, in turn, is the *formal cause*, which determines what external events can affect a system in principle. The processes that can change a whole, in turn, are the *efficient causes*. But they can cause changes only if that whole can potentially be changed by the given efficient cause. That is, what is being affected determines what can affect it and how it can be affected in principle. Consequently, efficient causality cannot be understood unless material and formal causes are understood at the same time as well. Final cause is as important as the other causes. It determines what can be the result of the change of the whole (for a thorough analysis of different theories of causality, see Toomela, 2019).

## Methods do not yet make methodology

Structural-systemic approaches also require that psychology develops a theory and philosophy of its scientific methods—a general *methodology* (for outlines, see Toomela, 2022). Mainstream psychology generally lacks an elaborated methodology. The common recipe-style books compiling ready-to-use methods, as used in quantitative psychology, do not yet make a methodology. Methodology, as the science of methods, explains how selected methods allow us to answer specific research questions. Each new question may require novel, methodologically grounded methods.

But many quantitative methods used in psychology (e.g., statistical tests) are not grounded in an elaborated methodology. They provide only probability statements but no theoretical justification about how these methods could allow us to address specific research questions and to explore specific study phenomena (Toomela, 2011, 2014b, 2022; Toomela and Valisiner, 2010; Uher, 2025; Valsiner, 2017). Importantly, such methods do not enable us to develop a structural-systemic understanding of psychical phenomena. Why?

Quantitative psychology largely studies only *observable* behavioural performances (e.g., test results, responses to questionnaires) while aiming to explore the *non-observable* psychical processes enabling them (e.g., intellectual abilities). However, *observably identical* behaviours may emerge from interactions of *different underlying* psychical processes (Richters, 2021; Sato et al., 2009; Toomela, 2007a, 2008b, 2009b; Uher, 2022b). But when observations are encoded into variables, such that *observably identical* behaviours are taken to arise from *psychically identical* processes, then the most important information is already lost because there is no way to discover what different processes may underlie observably identical behaviours (Toomela, 2008b; also Danziger and Dzinas, 1997; Maraun and Halpin, 2008; Uher, 2021a). For example, individuals can generate correct answers to simple arithmetical tasks by mentally calculating, counting their fingers or just recalling memorised answers. But which of these processes they have actually used remains unknown when only their responses are encoded. Psychological research that ignores this crucial point is, in fact, a version of behaviourism and thus, unable to explore psychical phenomena (Toomela, 2000, 2008a,b,c, 2011, 2014c, 2019).

The structural-systemic conceptualisation of the psyche as a complex system also highlights that, as structural wholes, psychical phenomena cannot be explored by reducing them to parts and studying these in isolation. Such *reductionist approaches* are commonly pursued in quantitative psychology, however, where wholes, described in constructs (e.g., 'intelligence'), are (conceptually) dissected into parts (e.g., verbal, numerical, spatial or reasoning abilities). Results obtained on these (conceptually) separated parts (e.g., different tasks in 'intelligence tests') are then simply combined (e.g., averaged), assuming the index score could be a 'measure' of the whole. Functional performances in higher cognitive abilities, however, are impossible without the involvement of various further processes (e.g., perception, reading comprehension ability, long-term memory). These must be present as well for complex cognitive processes to emerge at all. An individual's low performance in specific tasks therefore does not mean that the specific cognitive processes at which these tasks are targeted were not involved. Rather, it indicates only the individual's reduced or failed ability to use these processes in the given task situation (e.g., social pressure, noise). That is, complex cognitive processes can emerge only in the context of countless other concurrent processes and phenomena both internal (and thus, likewise hidden) and external to the individual (e.g., psychical, physiological, situational). This makes it impossible to determine the specific contribution that selected cognitive processes may make to observable task performances (Toomela, 2008b; Uher, 2022a, 2025).

These fundamental relations are elaborated also in another non-reductive ontology that focusses holistically on individual persons in the social, cultural and societal contexts of their lives. This *person-based ontology* (Martin, 2022) conceptualises human individuals as persons who are, at once, bio-physical and socio-cultural beings. Its origin can be traced back to Aristotle who conceptualised the human being as a bio-physical entity that develops within societies as a social and political being, thereby acquiring intellectual abilities (e.g., reasoning) and character (e.g., virtues). These *non-dualistic ontologies* differ profoundly from *Descartes' dualistic ontology* in which persons' material bodies are separated from their immaterial 'minds', which raises the fundamental problem of how these might interact, such as to enable action (body–mind problem). Descartes' dualistic ontology dominated Anglo-American philosophy and psychology, which also pursued reductionist approaches, in which persons are reduced to their bio-physical, behavioural and psychical parts, while their complex life contexts are reduced to quasi-laboratory settings and psychometric testing conditions (Martin, 2022). Conceptualising persons as ontological units, by contrast, allows for considering the inherent contextuality of psychical phenomena as well as for pinpointing the implications that this has for quantitative investigations.

## The contextual constitution of psychological phenomena does not yield to methods of quantitative measurement and laboratory experimentation

In his line of research on the psychology of personhood, Jack Martin has highlighted the idea that psychological phenomena are constituted by human interactivity within the life contexts of human beings (Martin, 2013, 2024; Martin and Bickhard, 2013; Sugarman and Martin, 2020). What interests us most in our everyday lives is neither accessible through nor reducible to bio-physical phenomena, which are amenable to precise quantitative measurement. Phenomena, such as identity, self-other understanding, perspective-taking, imagination, purpose, creativity or existential concern, are socio-culturally, historically and biographically constituted (Kirschner and Martin, 2010). The contextual constitution of psychological phenomena cannot be illuminated by methods of quantitative measurement and laboratory experimentation that have proven so successful in natural, bio-physical science.

### The socio-cultural life contexts of people

Our historically established socio-cultural communities are replete with practices, customs, traditions and ways of interacting, communicating and living. Our embeddedness and participation—from birth to death—within these contexts constitutes us as persons with self-other understanding, practical know-how, personal and collective identity, biographical storylines as well as moral and rational agency (Martin and Sugarman, 1999;

Martin et al., 2003, 2010). These contexts, and our interactivity within them, make up our *personhood* in ways that do not lend themselves to experimental variation in laboratory study or to standardised, quantitative measurement. Our personal and collective being and living initiate us into the possibilities and constraints afforded by our socio-cultural contexts (Danziger, 1990; Martin and Sugarman, 1999; Valsiner, 1998). Yet, in the course of our lives, we are able to develop ways of acting and interacting that alter these contexts. We humans are caught up in a circle of existence within which generations of us inherit, transmit and modify our life contexts during our own lifetimes.

## Problems of quantitative measurement and laboratory experimentation in psychology

Physical measurements in daily life and in science rely on *standard units of measurement*. Psychological measures, by contrast, rely primarily on *ratings* and *counts*. We have no objective, standard units with which to measure thoughts, ideas, opinions, emotions, actions, intentions, meanings or experiences—let alone to capture the more macro-level phenomena of human life, such as moral and existential concern that arise within the circles of existence that we inherit, adapt and pass on. Ratings of degrees of confidence, strengths of beliefs or levels of self-determination rely on the subjective judgements of researchers and research participants (Martin and McLellan, 2013; Uher, 2018a, 2022b, 2023a). Counting kinds of thought, frequencies of emotional occurrences or particular imaginings is unlike counting numbers of birds, heartbeats or users of public transit. Measuring physical states or processes is not akin to interpreting psychological states or processes (Lamiell, 2019b; Martin and Sugarman, 2009; Martin et al., 2015; Smedslund, 2021).

Unlike the trigonometry and calculus that can be applied to physics, psychology's statistical procedures do not enable precise point predictions and replications. Laboratory contexts in physics are specially constructed spaces for the careful observation and measurement of isolated phenomena under controlled conditions. Laboratory contexts in psychology, by contrast, mostly reduce and distort the everyday phenomena that they purport to study and 'measure'. The phenomena studied in psychological laboratories are literally and figuratively "out of context". In consequence, there is a large gap between the empirical findings of experimental psychology and the lives that we lead as historically situated, socio-cultural and biographical beings (Danziger, 1985a, 1990; Gergen, 2001; Martin, 2022, 2024; Valsiner, 2014a).

## Psychology as a socio-cultural practice and its impact on society

Psychological science is itself a multifaceted set of historically established, socio-cultural practices that affect us in somewhat predictable but sometimes also highly unpredictable ways (Martin, 2024; Valsiner, 2012). More than any other social science, psychology claims to foster factual and progressive understanding of our existence, actions and experiences. Such claims, however, require a better footing than that provided by much of the current experimental and professional psychology. Specifically, they require a reimagining that goes well beyond what some regard as a crisis of replicability in psychological research findings.

Since at least the mid-1980s, a growing number of social scientists and other scholars have become less interested in psychology and psychotherapy as purportedly applied sciences. They argue that, by trying to align psychology's scientific aspirations and status to those of physics and by focussing just on the efficacy of its professional practices, we risk missing out on the larger and arguably more important impact that psychological and psychotherapeutic ideas and practices can have on contemporary cultures, societies and individuals (Martin and McLellan, 2013; Madsen, 2014). Scholarly inquiry that examines connections between the lives, works and sociocultural impact of psychologists can provide valuable information about how psychologists and psychology affect people and their life contexts and experiences (Martin, 2017; also Fleck, 1935/1979).

## Alternative methods of psychological inquiry lead to new knowledge

Methods of life study, interpretation and writing, such as historical ontology (Hacking, 2002), biography and psychobiography (Kirschenbaum, 2007), ethnography (Rogoff, 2011), narrative inquiry (Hammack and Josselson, 2021), positioning theory (Harré and Van Langenhove, 1999) and life positioning analysis (Martin, 2013, 2024) aim to reveal dynamic reciprocities and relationalities that exist among people and their life 'worlds'. Such research can suggest possibilities for balancing conflicting demands for change and stability that attend the ongoing, mutual co-constitution of ourselves and our societies within contemporary life. In view of the ongoing social conflicts, complex real-world problems and the many crises in our societies, democracies and global relationships, such approaches have become more important than ever. A psychology of persons and of their lives must attend directly to their life concerns as these are experienced and lived—rather than as simulated and probed in comparatively decontextualised experimental settings with equally decontextualised pseudo-'measures'. Only by focusing on the actual lives and life conditions of real people can we, as psychologists, recognise and face directly the possibilities that we create for both humanity's flourishing and its peril with the aim of enriching the former and guarding against the latter (Martin, 2022, 2024).

This person-based ontology aligns with many of the ontological and epistemological commitments of critical realism. Different variants of realism and other philosophical theories have been developed in the sciences, many of which are also used in quantitative psychology. We discuss some of these now in our next Topic 2 with regard to the philosophy-of-science fundamentals underlying theories, methods and practices of psychological 'measurement'.

## Topic 2: Fundamentals of psychological 'measurement' and quantitative psychology—Crucial differences to genuine measurement

Physical measurement procedures are clearly not applicable in psychology given the peculiarities of its objects of research, such as their contextuality, developmental and socio-cultural constitution and inherent structural-systemic complexity (see Topic 1). Quantitative psychologists therefore developed their own definitions, concepts, theories and methods of 'measurement' (therefore here put in inverted commas) largely independently from those of measurement established in physical science and metrology (the science of physical measurement and its application; Berglund, 2012; Mari et al., 2021; McGrane, 2015; Uher, 2020a). Still, quantitative psychologists often draw analogies to physical measurement and interpret their findings as 'measurement' results that provide quantitative information about the phenomena studied in individuals. This entails conceptual errors because many psychologists are unaware of crucial differences in the underlying philosophies of science—and therefore also of contradictions that their conflation entails. Here we do not aim to provide a comprehensive comparison (see Uher, 2020a, 2021a,b, 2025). But we discuss key problems and important differences that are still largely overlooked. We exemplify these by specific theories and practices of psychological 'measurement'.

### Psychology's operationalist definition of 'measurement' and quantitative data generation with rating 'scales'

In her transdisciplinary line of research, Jana Uher explored theories, concepts and approaches of measurement and quantification across different empirical sciences. *Transdisciplinarity* gained recognition as a new way of thinking about and engaging in scientific inquiry since the 1970s. Unlike cross-, multi- and inter-disciplinarity, it is aimed at exploring complex systems and complex ("wicked") real-world problems that require the expertise of many scientific disciplines. Collaboration and integration across the sciences, however, are often hindered by discipline-specific jargon, theories, methods and practices. Transdisciplinarity[2] is therefore aimed at exposing disciplinary boundaries and the fundamental, often unstated beliefs on which scientific systems are built (presuppositions). Making these explicit is necessary to understand the non-obvious differences in discipline-specific processes of scientific inquiry—especially in their underlying ontologies, epistemologies and methodologies—as well as in their resulting bodies of knowledge. This also allows for discovering hidden connections between different disciplines as well as for generating unitary intellectual frameworks that rely on but also integrate and transcend different disciplinary paradigms (Bernstein, 2015; Gibbs and Beavis, 2020; Montuori, 2008; Nicolescu, 2008; Piaget, 1972; Uher, 2018a,b,c, 2021c, 2024, 2025).

Using transdisciplinary approaches, Uher analysed epistemological and methodological fundamentals of theories, methods and practices of measurement and quantification established in psychology, social sciences, behavioural biology, physics and metrology. Her analyses pinpointed commonalities and differences, especially between psychological 'measurement' (e.g., psychometrics) and physical measurement that are still hardly considered in pertinent debates (e.g., Uher, 2018a, 2019, 2020a, 2022a,b, 2025).

### What actually is quantity?

The most basic concept for quantitative sciences is that of *quantity*. Surprisingly, however, most scholars seem to rely on their intuitive understanding of quantity rather than a scientific definition. This entails confusion as to what measurement actually is, especially when mere categorisation is misleadingly termed 'nominal measurement' in psychology (Stevens, 1946) but also in engineering and metrology (Finkelstein, 2003; White, 2011). Some contend that measurement is solely defined through its process structure rather than also through a feature of its results (Mari et al., 2013). However, an elaborated process structure coordinating observations of the objects of research with our concepts, theories and models about them is basic to any form of elaborated scientific inquiry (Uher, 2025).

Ontological philosophy provides clear definitions. *Qualities* are properties that differ in kind (Latin *qualis* for "of what sort"). Length, weight, temporal duration and sound intensity are qualitatively different. *Quantities* (from Latin *quantus* for "how much, how many"), in turn, are divisible properties of entities of the same kind—thus, of the *same quality* (Hartmann, 1964). When qualitatively homogeneous entities change in quantity, such as by adding or dividing them, their meaning as entities of that specific quality remains unchanged. Placing several boxes side-by-side (concatenation) changes the quantity of their joint width but does not alter its quality as being that of length. That is, entities of equal (homogeneous) quality can be compared with one another in their divisible—quantitative—properties in terms of their order, distance, ratio and further relations as specified in the *axioms of quantity* (e.g., equality, ordering, additivity; Hölder, 1901[3]; Barrett, 2003; Michell, 1990; Uher, 2022a).

This highlights that *measurement has advantages over mere categorisation* by *additionally* enabling the descriptive differentiation between divisible instances of the same kind (quality)—between quantities (Hartmann, 1964; Michell, 2012; Uher, 2021c,d, 2022a). In this way, measurement enables more sophisticated analyses of categorised objects and their

2 There are two schools of transdisciplinarity. The present analyses build on *theoretical transdisciplinarity*. *Applied (practical) transdisciplinarity*, by contrast, is aimed less at developing theoretical frameworks and new forms of knowledge but more at understanding real-world problems and developing tangible solutions. It involves scholars from different disciplines but also political, social and economic actors as well as ordinary citizens with the aim of producing socially robust knowledge rather than merely reliable scientific knowledge (Uher, 2024).

3 For an English translation (see Michell and Ernst, 1996, 1997).

relations. But it requires appropriate qualitative categorisation of study phenomena, which is far more challenging than the identification of divisible properties in them. Yet both may also go hand in hand as the history of metrology shows (e.g., development of thermometers; Chang, 2004). Hence, ultimately, *all quantitative research has a qualitative grounding* (Campbell, 1974; Kaplan, 1964). The common dichotomisation of 'quantitative' vs. 'qualitative' methods, data and approaches reflects a fundamental misconception, implying quantities could be determined independently of the quality studied, yet overlooking that any quantity is always *of something*—a specific quality (Uher, 2018a, 2020a, 2022b, 2023a).

## Steven's redefinition of 'measurement' and concepts of 'scale' types

Obviously, psychical phenomena lack properties that are amenable to concatenation, thus failing to satisfy the *additivity* criterion of quantity (Ferguson et al., 1940). To establish quantitative inquiry in psychology regardless, Stevens (1946) proposed that psychologists should focus not on properties featuring demonstratively additive structures but instead on the structure of the operational procedures that are used for empirical inquiry (Borsboom and Scholten, 2008). For this purpose, he turned to *operationalism* from physics (Bridgman, 1927) and adapted it in his own specific ways (Feest, 2005) by claiming.

> "operationism consists simply in referring any concept for its definition to the concrete operations by which knowledge of the thing in question is had" (Stevens, 1935, p. 323).

In line with this, Stevens (1946, p. 667) *defined* '*measurement*' as "the assignment of numerals to objects according to a rule". This operationalist redefinition formed the basis for psychology's theories and practices of 'measurement' and separated them from those of measurement used in physics and metrology (Mari et al., 2021; McGrane, 2015; Uher, 2020a, 2021a, 2025).

Many psychologists seem to be aware neither of how fundamental the thus-introduced differences are nor of the epistemological errors on which these are built and that these entailed. For example, Stevens' redefinition promoted the idea that psychology requires a *"soft", "weak" or "wide" definition of measurement* (Eronen, 2024; Finkelstein, 2003; Mari et al., 2015). Certainly, psychology does not need the high levels of measurement accuracy and precision that are necessary for sciences like physics, chemistry and medicine where errors can lead to the collapse of buildings, chemical explosions or drug overdoses (Uher, 2023a). But simply redefining a scientific activity that is as fundamental to empirical science as measurement is *epistemologically mistaken* because this undermines its comparability across the sciences. Specifically, redefining measurement for non-physical sciences fails to provide guiding principles that specify how genuine analogues can be conceptualised and empirically implemented while appropriately considering the peculiarities of the different sciences' study phenomena. Epistemic comparability is crucial for research on complex real-world problems because integrating findings across different sciences presupposes transparency in their quantitative data generation to enable epistemically valid inferences on the phenomena studied. Labelling disparate procedures uniformly as 'measurement' also obscures essential and necessary differences in theories and practices between the different sciences as well as inevitable limitations (Uher, 2022a, 2025).

Indeed, following Stevens' redefinition, many psychologists came to understand 'measurement' as simply any consistent operational procedure of numerical assignment (McGrane, 2015). Many psychologists also know only Stevens' (1946) concepts of *'measurement scales'* (nominal, ordinal, interval, ratio)—which likewise depend on operational rules of numerical assignment (Borsboom and Scholten, 2008)—ignoring that these are neither exhaustive nor universally accepted (Thomas, 2019; Uher, 2022a; Velleman and Wilkinson, 1993). Stevens' operationalist approaches offered simple solutions for enabling empirical research and theory development in quantitative psychology. Still today, operationalism is considered an essential feature of rigorous psychological research, where constructs are defined through operational procedures, such as ratings on sets of item statements describing the phenomena of interest (AERA et al., 2014).

Stevens' works also informed one of the first theories of 'measurement' established in the social sciences as well as psychology's main method for generating quantitative data.

## Representational Theory of Measurement (RTM)

Representational Theory of Measurement (RTM) formalises axiomatic conditions by which relational structures observable in an object of research can be mapped onto relational structures in a symbolic system (e.g., model with variables and numerical values). It provides mathematical theories for this mapping (*representation theorem*), including permissible operations for transforming the symbolic relational structures without breaking their mapping relations onto the empirical relational system studied (*uniqueness theorem*; Krantz et al., 1971; Luce et al., 1990; Narens, 2002; Suppes et al., 1989; Vessonen, 2017). The theory's focus on *isomorphisms*—thus, on reversible one–to–one relations between observables and numerical data—presupposes that the objects of research feature properties with quantitative relations that are directly observable (e.g., 'greater than' or 'less than'). Such relations can be mapped straightforwardly onto a symbolic system that preserves these relations (e.g., ordinal variables; Suppes and Zinnes, 1963).

In psychology's complex study phenomena, however, quantitative properties obviously cannot be identified—the very fact that first led Stevens to focus instead on operational procedures. Psychologists therefore relied on Stevens' concepts of 'measurement scales', which define types of data variables by their formal properties (e.g., ordering relations, equal distances), thus specifying also the formal transformations (e.g., arithmetic operations) that can be performed in the symbolic relational system. Following the isomorphic relations between the empirical (real) and the symbolic (formal) system stipulated by representational theory—as well as the 'measurement' jargon used—these merely formal concepts were also ascribed the meaning of 'instruments', analogous to physical measuring devices. Physical measuring instruments (e.g., weighing scales) enable

traceable empirical interactions with the *specific quantity to be measured* (the *measurand*; e.g., an apple's specific weight). Instrument, measurand and their empirical interaction are all physical and pertain to the real system under study, whereas the information about them is symbolically encoded in the formal study system (e.g., model with variables and values).

In psychology, however, these crucial epistemic distinctions are obscured because psychological 'instruments' are language-based—and thus, formal as well (see Topic 3; for details, Uher, 2025). For example, the term *psychometric 'scales'* is used to denote the items and answer 'scales' (e.g., five answer categories) presented to respondents (e.g., digitally) as 'instruments' that are thought to enable interactions with the study phenomena (e.g., respondents' beliefs). In this notion, they pertain to the real study system. But the term also denotes the statistically modelled (latent) structures underlying the response values obtained on many (manifest) item variables (modelling, e.g., probabilistic response patterns). In this second notion, psychometric 'scales' form part of the formal study system—respondents neither know about nor interact with it (for details, Uher, 2025). Referring to 'scales' indiscriminately as parts of both the empirical *and* the symbolic relational systems obscures the crucial epistemic distinction between them (Uher, 2018a, 2022b). This also disables the epistemic necessity to specify the relations between them.

The relations between real (empirical) and formal (symbolic) study systems concern one of the most fundamental problems in empirical science. Their specification requires *representation decisions* about what to represent, and what not, and about how to represent this in a formal system (e.g., a model; Harvard and Winsberg, 2022; Uher, 2025). This is discussed as the *problem of scientific representation* in philosophy of science (Frigg and Nguyen, 2021; van Fraassen, 2008), as *encoding and decoding relations* in biophysics and theoretical biology (Rosen, 1985, 1991), as the *problem of coordination* or *correspondence* in physics (Hempel, 1952; Margenau, 1950), and as *coordination and calibration* in metrology (Chang, 2004; Luchetti, 2020; Tal, 2020). Many psychologists, however, seem largely oblivious of these fundamental issues. Some even consider representation as irrelevant for psychological 'measurement' (e.g., Borsboom and Mellenbergh, 2004; Michell, 1999)—a consequence and reflection of Stevens' operationalism. Indeed, neither Stevens nor representational theory provide any concepts or procedures for how and why some empirical observations should be mapped to a symbolic relational system (Mari et al., 2017; Schwager, 1991). Rather, they stipulate purely representationalist and operationist procedures focussed solely on the assignment of numerical values with mathematically useful relations.

## Quantitative data generation with rating 'scales'

These procedures also underlie psychology's primary method of quantitative data generation—rating 'scales'—in which numerals (e.g., '1', '2', '3', '4', '5') are rigidly assigned to the answer categories provided to raters (e.g., five stages indicating levels of agreement). Misled by the premises of an efficient implementation of 'measurement' in psychology, many overlook even striking errors in their own numerical assignments. Indeed, what justifies the assumption that 'agree' (assigned '4') reflects more than 'disagree' (assigned '2')? Is agreeing with something not rather an entirely different idea than disagreeing with it? How can we assume that 'neither disagree, nor agree' (assigned '3')—thus, having no opinion or finding the item not applicable—constitutes more than 'disagree' (assigned '2')? And why should we assume that the distance between 'neither disagree, nor agree' and 'agree' equals that between 'agree' and 'strongly agree' (both assigned a distance value of '1')? Given the logico-semantic meanings of these verbal answer categories, it is unsurprising that raters interpret them not as reflecting order or even interval relations but only as categorically—thus, qualitatively (nominally)—different. Such logical errors also occur with frequency 'scales'. Given that occurrence rates generally differ between phenomena (e.g., chatting vs. arguing), rating 'scales' force raters to indicate a broad range of quantities *flexibly* in the same bounded answer 'scale'. Raters can do so only by assigning *different* quantitative meanings to the *same* answer value—a necessity that violates core ideas of measurement (Uher, 2018a, 2022a, 2023a, 2025).

Nevertheless, rating data are commonly interpreted as results of 'measurement' that provide quantitative information about the phenomena of interest (e.g., individuals' beliefs). This contrasts with their purely operationalist generation in which numerical values are assigned to the fixed 'scale' categories in identical ways for *all* items of a questionnaire, *regardless of the specific study phenomena* to which these may refer. That is, without explanation, raters' judgements of verbal statements, such as their levels of agreement, are re-interpreted as reflecting quantities of the phenomena described. Many psychologists seem to be unaware that this interpretation involves a shift in the underlying philosophy of science because psychological theories and practices build on different presuppositions than the measurement framework established in physics and metrology (Uher, 2020a, 2021a, 2025).

## Confusion of two incompatible philosophies of science masked by psychological 'measurement' jargon

Stevens' operationalism and representational theory of measurement are strongly connected to *positivism*, coined in particular by Comte in the 19th century for social science. This family of philosophical theories builds on the presupposition that scientific knowledge should be derived solely from empirical evidence of observable phenomena. Inspired by the successes of the natural sciences, positivists seek to provide accurate and unambiguous knowledge of the 'world', thought to be objectively given and independent from us, using natural science methods—observation, experimentation, logic and mathematics. Scientists' tasks are to study the facts (thus, focussing on the concrete), to identify regularities in them (therefore focussing on replicability) and to formalise these in (descriptive) laws, whereby explanations often involve no more than subsuming special cases (particulars) under general laws (see Topic 4). Positivists reject abstract theorisation and metaphysical beliefs, which are dismissed as speculative, unobservable and untestable. *Metaphysics*[4], dating

4 The term 'metaphysics' has a history of various meanings. Originally, it indicated only the order of Aristotle's works, in which it happened to be listed after those written on physics (Ancient Greek *meta* meant 'after'). It is also

back to Aristotle, is the philosophical inquiry into abstract principles and the first causes of things, covering topics such as ontology (being), space, time, determinism and free will (van Inwagen et al., 2023). The positivists' view that eliminating metaphysics would be desirable, however, is a metaphysical presupposition itself (Bickhard, 2001). Hence, positivism is focussed on description, control and prediction (replicability)—yet at the expense of advancing an ontology of the objects of research and their nature, which limits its ability to develop explanations of them (Al-Ababneh, 2020; Ali, 2023; Howell, 2013).

Physical measurement, by contrast, builds on theories of realism (Mari et al., 2021; Schrödinger, 1964; von Neumann, 1955; Uher, 2025). *Realism* generally is the philosophical perspective that there is a 'reality' that exists regardless and independently of our perceptions, understanding and beliefs of it. This requires ontological theories about the objects of research and epistemological theories about the ways in which knowledge about these objects can be gained. This general perspective underlies many different forms of realism, each involving different epistemologies (e.g., scientific realism, critical realism) and used in different variants, often reflecting their authors' idiosyncratic qualifications. We do not aim to provide an overview here but select only some that are relevant for our analyses.

Theories of *scientific realism*, for example, involve the presuppositions that both observable and not directly observable parts of 'reality' exist (e.g., electrons) and that we can explore these with our best scientific theories and models—thus, using both empirical observation and theoretical reasoning. The main epistemic belief is that science aims at providing an accurate, truthful account of 'reality' so that, with scientific progress, accepted theories are believed to approximate that 'reality' ever more closely. Specifically, theories are regarded as truthful to the extent that their concepts correspond to the real study system, which underlies the successful use of these concepts for advancing theoretical explanations of these real systems (Chakravartty, 2017; Miller, 2024; Al-Ababneh, 2020).

This pinpoints key differences to positivism where theories are aimed only at describing and predicting observable phenomena, as evident in many quantitative psychologists' focus on replicability, predictive validity and other common quality criteria of mainstream psychology. Therefore, psychometricians who (implicitly) rely on positivist presuppositions often are simply "not persuaded" by the necessity to establish theoretical and empirical relations between the real and the formal study system. They also often refer to realist theories as "axiomatic measurement theory", implying a metaphysical notion (e.g., Markus, 2021). Yet without systematically conceptualising and empirically connecting the real and the formal study system, results cannot be interpreted as reflecting quantitative information about the phenomena studied in individuals (Rosen, 1985). This lack of epistemic validity contradicts the psychometricians claim to be able to "measure the mind" (e.g., Borsboom, 2005) as well as calls to consider ontological theories in psychological 'measurement' (Borsboom, 2006). This (implicit) reliance on two incompatible philosophies of science—one for the theories and empirical practices, and another for the result interpretations and declared aims—causes logical contradictions (Uher, 2020a, 2021a,b, 2022a, 2025).

The correspondence between theoretical concepts and empirical observations is central to the *problem of universals*—identified already by Plato, Aristotle and scholars of the medieval university. It concerns the fundamental epistemological question of how we can develop universal categories and trusted knowledge of nature if we can always observe only a finite number of concrete particulars (Klima, 2022). Over millennia, scholars developed many approaches to explore this problem. Our next contribution acknowledges the constructed nature of theoretical concepts and their pragmatic utility while simultaneously endeavouring to establish a systematic mapping to the empirical study system. These presumptions are used to explore theoretical concepts and models of psychological 'measurement' and to pinpoint the contradictions that are still not well considered.

## Measurement in psychology: A promise that failed to materialise

Psychology's efforts to establish a robust system for measurement have faced profound conceptual, theoretical and methodological challenges. Since the early days of scientific psychology, there has been a tendency to develop 'measurement' models that are mimicking those used in classical physics (Heene, 2011; Cornejo and Valsiner, 2021). This approach was intended to call for a "natural science infinitely more complete than the psychologies we now possess" (James, 1895, p.124)—thus, for the naturalisation of psychological science. It has become evident, however, that this enterprise has failed. In their measurement-theoretical research, Jörg-Henrik Heine and Moritz Heene highlighted that the most basic approach to measurement involves the simple principle of *counting* units. This requires that a one-to-one relationship is established between the phenomenological object domain and the mathematical metric space of positive real numbers (Heine and Heene, 2025)—the most basic approach to measurement (von Helmholtz, 1887; Hölder, 1901). However, this has never been successfully applied in psychology's entire history.

Why did the promise of metric measurement in psychology remain unfulfilled? Heine and Heene's (2025) critique of the one-sided focus of psychometric models on the numerical relational system highlighted various conceptual, theoretical and methodological issues. These issues cast a merciless light on the deep gap between mathematical models for $\Theta$ and the empirical relational system $\Psi$.

### Conceptual issues: Misconstrued operationalism and jingle–jangle fallacies

Conceptual issues arise from the inherent complexity of psychological constructs and the empirical problems that this entails (Maraun, 1998). Unlike the natural sciences, where technical concepts are clearly defined and applied by necessary rules, psychological constructs are rooted in everyday language. *Operationalism* (Bridgman, 1927, 1938)—as used in psychology

often misinterpreted as denoting 'what goes beyond physics or reality', linking it to speculation.

to bridge the gap between Θ and Ψ (Feest, 2005)—has instead deepened it. Originally intended as an "operational analysis" (Bridgman, 1938) to explicate "the meaning-contours of concepts *already in place*" (Koch, 1992, p. 261, emphasis in the original), psychology misconstrued operationalism as a framework for *defining* constructs by naming its (purportedly) quantifiable entities (Koch, 1992; Chang, 1995; Hibberd, 2019). Therefore, psychological 'measurement' often relies on *nomic measurement* (Chang, 1995, p. 153), whereby unobservable constructs are linked to observable proxies through a-priori definitions and settings. For this reason, it is also called *measurement by fiat*—'measurement' by decree (Torgerson, 1958, p. 22; Cicourel, 1964, p. 3; Uher, 2020a). This, however, can lead to circular reasoning (van Fraassen, 2008; Chang, 2004, 1995; Luchetti, 2024; Uher, 2021a, 2025). This operationalist practice also resulted in a plethora of *different* 'definitions' of constructs sharing the *same* term that frequently show only empirically weak correlations with each other (Elson et al., 2023; Pace and Brannick, 2010; Skinner, 1996). It also led, vice versa, to the proliferation of *different* terms for the *same* construct—thus, contributing further to *jingle-jangle fallacies* (Hanfstingl et al., 2024; Kelley, 1927; Thorndike, 1903).

## Theoretical issues: Fragmented theories and misguided assumptions about measurement and replicability

Psychology is currently debating Questionable Research Practices (QRPs) as potential causes of its replication crisis. But psychology still lacks robust discussions about the Questionable Research Fundamentals (QRFs) of its 'measurement' concepts, such as the near-exclusive reliance on continuous variable models to explain abstract population-level effects through aggregate statistics (Figure 1). This (still largely) unquestioned practice reflects the widespread misuse of ergodic assumptions, where intra-individual and inter-individual variations are treated as equivalent (see Topic 4; Molenaar, 2008; Speelman et al., 2024). Such an assumption fails to account for the idiographic and developmental nature of psychological processes, where individual differences are crucial (Salvatore and Valsiner, 2010). When unaddressed, this oversight can contribute substantially to psychology's replication crisis.

Psychological research should instead emphasise empirically observable patterns and structures to uncover the underlying idiosyncratic mechanisms and causes of its study phenomena, aligned with an "observation-oriented science" approach (Grice et al., 2012). Some psychological theories, however, have been criticised for inspiring empirical research on hypotheses that are trivial or logically self-evident, thus offering little value to scientific understanding. For example, Bandura's (1977) self-efficacy theory was identified by Smedslund (1978) as a starting point for pseudo-empirical follow-up research. Smedslund demonstrated that the core propositions of the theory could be reformulated into 36 a-priori, non-contingent theorems—thus, statements that are logically provable without requiring empirical validation (see also Smedslund, 1988, 1991, 2016). The motivation for some of these pseudo-empirical research projects may lie in a simplistic logic of justification, which is often seen in the context of educational policy decisions.

Another fundamental issue in psychology is that researchers frequently compare empirical outcomes with one another instead of testing them against a theory to be validated or refuted (Muthukrishna and Henrich, 2019). The Reproducibility Project (Open Science Collaboration, 2015) illustrates this dynamic. As the former NASA scientist Paul Lutus (personal communication with Moritz Heene, 3rd March 2016) put it:

> "the Reproducibility Project can be carried out with predictable consequences, then many people will discuss the outcome in great detail without anyone noticing that *the root problem in psychology is that investigators are comparing experimental outcomes with each other, rather than with a theory to be either supported or falsified.* Modern psychology is an intellectual construct in which everything lies at the periphery, but there's nothing at the centre to bind the periphery together. In psychology, and if it were possible, that centre would be a robust theory against which every experiment would be compared, and either a problem with the experiment would be revealed or the theory would be modified or discarded, replaced by a better one, as regularly happens in physics" (italics added).

Unlike physics, where theories (e.g., Newtonian mechanics) provide a foundation for measurement, psychology's reliance on fragmented constructs and study phenomena hampers the integration of its large empirical databases into a cohesive scientific framework (Michell, 2000).

## Methodological issues: Misapplying natural science paradigms to psychology

Methodologically, an over-reliance on natural science paradigms (naturalisation; Sherry, 2011) has resulted in the use of inappropriate analogies for 'scales' (Stevens, 1946, 1958). This theoretical gap between the mathematical models and the empirical psychological 'reality' is further highlighted by the limitations of psychometric models, such as Rasch models (Rasch, 1960). Heine and Heene (2025) criticised the widespread "putting-the-cart-before-the-horse" belief that relying on psychometric models merely as models for numerical relational systems could guarantee genuine interval-level measurements for psychological constructs. Early attempts were made to connect numerical and empirical relational systems (Fechner, 1858, 1860a,b). However, these efforts have been overshadowed by misinterpretations and misapplications of psychometric models—as if their mere application inherently yields interval scales for Ψ. In fact, their mere application generates real numbers for Θ while disregarding the relationship between the numerical and the empirical relational system, thereby potentially misrepresenting the true nature of psychological attributes (von Kries, 1882; Trendler, 2009; Uher, 2021a, 2025). Along those lines, Heine and Heene (2025) highlighted that the Rasch paradox is genuine: the 'interval scales' created by this item response model are consistent with nothing more than ordinal attributes of psychological variables (also Barrett, 2003; Michell, 2014; Trendler, 2022b). The same applies to conjoint measurement (Trendler, 2019a).

### Psychology's prospects for quantifying its study phenomena and future directions for novel developments in its methodology

The persistent challenges in measuring psychological phenomena originate from and are perpetuated by conceptual ambiguities, theoretical fragmentation and methodological misconstruals of (especially psychometric) models. Without addressing these fundamental issues, the promise of a robust and scientific measurement framework in psychology is unlikely to materialise. On the other hand, as Schönemann (1994, p. 150) suggested, we may need to accept

> "the prospect that psychology will never make much progress towards becoming a quantitative science" *in the sense of measurement in a metric space* also known as the real number line. Instead, psychological methodology must recognize that "… models can also be used that, from the outset, … imply only an ordinal scale level for both Θ and Ψ, such as the ordinal probability models" (Heine and Heene, 2025, p. 22).

The logical contradictions that arise from the positivist theories and empirical practices established in psychological 'measurement' and the realist interpretations of results and declared aims become obvious in further ways.

## Latent variable models, unit-free 'measurement' and generative artificial intelligence (genAI) cannot enable measurement: Psychology must consider the peculiarities of human mind

"Science requires measurement"—this belief has become quantitative psychology's unquestioned imperative (Michell, 2003). It builds on Thorndike's credo that everything exists in some amount and can thus be measured (Michell, 2020) as well as on Lord Kelvin's dictum that only what is expressed in numbers constitutes scientific knowledge (Barrett, 2005). In his research papers, blogs and postings, Paul Barrett expressed critical views on psychological 'measurement' that he had developed in his various roles not just in academic psychology but also in forensic psychiatry and the assessment industry. For 26 years, Paul Barrett maintained the IDANET[5] mailing list where he regularly informed a growing community of scholars about new publications of both mainstream and non-mainstream research and stimulated thought-provoking discussions on psychology's theories and practices of 'measurement'.

### The quantitative imperative and the myth of unit-free 'measurement' in psychology

Barrett (2005, 2008) advocated for rethinking the entire basis on which psychology generates its 'measurement' concepts. He highlighted that empirical experimental manipulations of attributes are required before attribute magnitudes can be represented by a real number system, let alone the instantiation of a unit of measurement. Without any empirical evidence suggesting that psychological attributes vary quantitatively (e.g., ordinal and additive structures), we should not make that assumption (Barrett, 2003, 2018; Michell and Ernst, 1996, 1997; Michell, 1997, 1999, 2000; Trendler, 2009, 2013). The most reasonable assumption is that we can assess partial orders or classes with some degree of 'fuzziness' between boundaries. Yet without any clear methodology for determining precisely how an attribute varies and what is causal for those variations, we are relying upon mere 'common-sense' judgements of magnitude (Barrett, 2018). Barrett also showed that neither unit-free 'measurement' nor arbitrary units can possibly sustain a quantity—whether trying to express it as a derived or a base unit[6] quantity (Barrett, 2011, 2018; Newell and Tiesinga, 2019). This was also elaborated upon by Trendler (2019b, 2022a) for psychological 'measurement' generally as well as specifically for conjoint measurement and Rasch modelling (Trendler, 2019a, 2022b), supported more recently by Heine and Heene (2025).

So, what remains of decades of research on latent variable models (LVM), hierarchical multilevel modelling (HML) and structured equation modelling (SEM)? Revelle's (2024) article "The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny" already answered this question in its title. The fundamental problems with psychology's unsupported assumptions of the human mind's measurability are not solved with ever more sophisticated statistical and visualisation techniques. 'Network psychometrics', for example, merely reifies as 'explanatory' what is essentially a simple network analysis and data visualisation application but hardly an advance in our understanding and explanation of the human mind. Barrett (2024) demonstrated (e.g., using computer simulations) that no psychometrics or test theory does more than provide general statements of 'effect' or 'measurement'.

### Most branches of mathematics are concerned with non-quantitative structures and provide meaningful concepts for formalisation in all sciences

Contrary to most psychologists' beliefs, studying psychological attributes and phenomena does not require quantitative 'measures' and not all structures studied with mathematics are quantitative. Mathematics is the science of abstract structure (Resnick, 1997). Most of its branches are therefore non-quantitative, such as pure mathematics, category theory, geometry or set theory, which provide important concepts for formalisation in empirical sciences (Barrett, 2003; Linkov, 2024; Parsons, 1990; Rudolph, 2013). Psychology requires non-quantitative 'measures'—classes, orders, structured observations and models (Barrett, 2003). These possess

5 Individual Differences and Assessment Network (IDANET); https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=idanet.

6 In metrology, base units are conventionally defined entities (e.g., metre, second) that are used as references for quantitative physical properties that cannot be expressed in terms of other quantitative physical properties (e.g., length, time). Derived units, by contrast, are conventional references derived from combinations of some base units (e.g., volume from length, velocity from length and time; Uher, 2020a).

a *pragmatic* value or are associated with "good enough" reasoning rather than any sophisticated statistical modelling or deployment of a statistical 'measurement' model.

It is convenient to use numbers to represent 'magnitudes' on occasion and to rely upon the arithmetic properties afforded by such use (e.g., in educational assessment). However, it must always be made clear that this numeration is solely for computational convenience rather than enabling any degree of accuracy of a 'measurement' of psychological attributes—for which Barrett (2003) proposed the term *applied numerics* instead of psycho-'metrics'.

## Generative AI and large language models (LLMs) cannot solve psychology's problems: Human minds are complex, open, self-organising systems

Barrett (2024) was an outspoken critic of the increasingly popular attempts in individual differences psychology to use Machine Learning (ML) algorithms for forming machine-generated 'measures' of personality attributes. He highlighted that, unlike physical scientists, we are not calibrating an alternative measure of length using a previously calibrated measuring instrument (e.g., a ruler or steel tape). Length can be formalised in a quantitatively structured base unit variable—but a personality attribute cannot. Consequently, attempting to predict personality scores with sufficient accuracy and generalisability such that they could be replaced by machine-generated scores using other kinds of observational or 'digital-footprint' data was never going to work—from the first principles of 'measurement' let alone the conceptual and known semantic haze of verbal 'scale' content (see Topic 3).

Human minds are not closed systems that can be manipulated and measured, as one might pursue with mechanistic variables in closed physical systems. Human minds are complex, open and self-organising cognitive systems (Barrett, 2005; Kelso, 1995; Trendler, 2009). Higher-order complex systems feature interconnected parts, non-linear dynamics, emergence, adaptation, sensitivity to initial conditions, feedback loops, equifinality and further peculiarities many of which are not found in inanimate systems (see Topic 1; Barrett, 2024; Uher, 2021c, 2025). The outputs of such systems cannot be accurately predicted, although they can be generalised and classified in terms of broad descriptive phenomenal statements. Many psychologists and technical people working on 'predictive' models seem to ignore, or be unaware of, the fundamental properties and qualities of the specific study systems whose outputs they are trying to model—that is, those of human beings.

## Alternative approaches that do justice to the individual: Observation-oriented modelling (OOM)

For causal analyses that do not rely upon assumption-laden statistical parameterisation and metaphorical discussions about 'unobservable variables', James Grice developed *Observation Oriented Modelling* (OOM; Grice, 2011; Grice et al., 2017a). As with actuarial analytics, the outcome is expressed as "how many cases actually showed the expected or hypothesised outcome? Was this by chance alone? And who were they?". Paul Barrett also showcased his own actuarial approach to these questions (Grice et al., 2017b; see Topic 4).

Quantitative psychologists cannot hide from their responsibility when, in courts (e.g., US Supreme Court), latent variable or average IQ scores, expressed to two-decimal place precision, are used (even if just partly) to make decisions relating to an offender's death penalty. In many countries, case-law has developed on the basis of popular beliefs about the epistemic authority of psychological 'measurement' although there is no empirical evidence that the IQ varies as a quantity or indeed as an equal interval attribute. It is just a matter of time until psychometric scores will be challenged in courts, as has previously occurred with forensic psychologists' and psychiatrists' diagnostic practices (Barrett, 2018; Faust, 2012).

The fundamental issues of psychological 'measurement' and the direct implications that they can have for individuals and society are increasingly discussed also in the public, such as with regard to high stakes testing, admission metrics and policies in educational and occupational assessment. Tackling these issues requires philosophical approaches that enable careful and epistemically justified interpretations of empirical findings.

## Realist philosophies of science for studying psychical and socio-cultural phenomena

The peculiarities of psychology's study phenomena (e.g., contextuality, socio-cultural constitution, higher-order complexity; see Topic 1) led to the development of further forms of realism. These involve epistemologies that are more appropriate for exploring individual (subjective) and socio-cultural (inter-subjective) interpretations, explanations and appraisals of observable and non-observable phenomena—that is, the meanings that these have for individuals and communities, psychology's central objects of research (Wundt, 1897; Uher, 2025). The existence of these meanings, as individual and socio-cultural phenomena, is conceptualised in realist ontologies. But non-realist epistemologies are used to consider that any scientific inquiry of such phenomena is always situated in a socio-cultural context that influences and shapes the process of inquiry. Moreover, all scientists are human beings themselves with their own personal and socio-cultural perspectives, contexts and frames of reference, which they bring (unwittingly) to their research. Therefore, psychologists cannot be independent of their study phenomena, which entails risks of unwittingly introducing pronounced ego-centric and ethno-centric biases into their research (Adam and Hanna, 2012; Danziger, 1990; Gergen, 1973; Faucheux, 1976; Uher, 2015a, 2020b, 2022b; Weber, 1949).

*Critical realism*, for example, builds on the presuppositions that the social 'world', just like the material 'world', features complex structures and that these exist independently of our knowledge of them. In social systems, observable phenomena can be explored for their underlying processes and causes (e.g., human agency). Critical realism emphasises the 'reality' of the study phenomena and their knowability but also that our knowledge about this 'reality' is created on the basis of our practical engagement with and collective interpretation and appraisal of that 'reality'. This allows for reflecting on the relation between the researcher and

the researched and for acknowledging that knowledge is theory-laden, socio-culturally embedded and historically contingent (see Topic 1). Hence, critical realism combines a realist ontology with a relativist epistemology, in which diverse perspectives (and even contradictions) are accepted, tolerated and valued (Bhaskar and Danermark, 2006).

*Constructivist realism* is another philosophical perspective that builds on the presuppositions that real-world phenomena (e.g., individuals' intellectual abilities) exist and that their narrated interpretation is intersubjectively constructed and negotiated in the context of their use. It highlights that formal models are human constructions (of analysts) that are used to represent important patterns of complex real-world phenomena in ways that suit the inferences intended. Models necessarily involve abstraction, simplification and idealisation and are studied, in applied work, regarding their aptness for a given purpose rather than simply their truthfulness. Therefore, model-based reasoning involves not just a dyadic relation between a model and real study system but a four-way relation among a model, a situation, a user and a purpose. That is, constructivist realism combines a realist ontology with a constructivist epistemology. It is used in our next contribution to explore meaning-making as a fundamental aspect of psychological 'measurement' in educational assessment, where it allows for considering multiple socio-cultural meanings of test results, models and applied practices (Kane, 1992; Messick, 1989; Mislevy, 2009, 2018).

## The contextuality of human experience and learning requires a socio-cognitive perspective on psychological inferences in educational assessment

Between-persons Latent Variable Models (LVMs[7]), such as those based on item response theory (IRT), trace back to trait psychology and were advanced, amongst others, through Spearman (1904). Despite their practical value in educational assessment (Lord, 2012), however, a widening gap exists between the LVM conceptualisation and the advances made in cognitive and social psychology to understand learning and acting—including performing in educational assessments. Robert Mislevy argued that a socio-cognitive perspective on LVMs can retain their pragmatic value, while avoiding conceptual errors inherent to current conceptions of LVMs (Mislevy, 2018, 2019, 2024).

### Latent variable models (LVMs): Key concepts and inherent problems

The kernel of LVMs is the function $f\left(x_{ij}|\theta_i, \beta_j\right)$. It formalises the probability density of a variable $x_{ij}$ for evaluated learner performances, given the latent ability variables $\theta_i$ of person $i$ and in task (item) $j$ characterised by parameters $\beta_j$ (e.g., difficulty). The common trait perspective invites taking $\theta$s as the persons' measures on a general psychological property $\Psi$, interpreted through a construct that is assumed to somehow cause the learners' performances $X$s. Conceptual errors often occur because assessment developers and users tend to conflate several distinct elements: the construct itself, the latent variable $\theta$ used to operationalise a person's ability, the underlying psychological properties $\Psi$ that the latent variable is intended to represent, and the observed assessment outcomes X. Importantly, LVMs are silent as to the psychological nature of $\theta$ and the socio-cognitive processes by which performances arise. Moreover, LVMs often fail to establish the measurement requirements that are necessary to epistemically demonstrate that the psychological property $\Psi$ intended to be studied does indeed exist.

### The socio-cognitive perspective on educational assessment

The socio-cognitive perspective synthesises research from psychology, linguistics, educational science and complex systems as to the nature of individuals' capabilities and how they develop these through interactions in their social milieu (Gee, 2021; Sperber, 1996). It conceptualises how individuals navigate through situations that are shaped by *linguistic, cultural and substantive regularities of knowledge and action*, which vary over times and contexts. Specifically, individual learners develop cognitive resources to recognise these regular patterns and to act through them. Although individuals are unique, interaction is enabled when individuals' experiences with respect to relevant linguistic, cultural and substantive patterns show similarities, leading to similar cognitive resources.

In any given assessment, individuals blend the particulars of the test situation with the cognitive resources that they have developed from previous experiences in their history of interactions in a cultural milieu. Educators' tasks are to identify linguistic, cultural and substantive patterns that are important for students' learning in order to develop suitable resources (curriculum), to provide the necessary learning experiences (instruction) and to obtain information about students' progress (assessment). By providing conceptual coherence, a socio-cognitive perspective helps to integrate instruction, assessment and real-world practices by explicating and leveraging linguistic, cultural and substantive patterns (Gee, 2008; Harris et al., 2016).

### Managing evidence, inference and argumentation in LVM-based assessments from a socio-cognitive perspective

This socio-cognitive perspective for assessment necessitates re-conceptions of educational 'measurement' and LVMs. While psychometric methods and concepts remain useful for differentiating between individuals' performances—from a socio-cognitive perspective—the focus shifts from 'measuring general psychological properties' to managing evidence, inference and argumentation for making such differentiations. Educational assessment still centres on a *construct* (Messick, 1995) but without being conflated with latent variables, general properties and measures. Here a construct is a *natural language concept*—what individuals can think or say, such as about what they do in situations. These constructs are conceived from a historical, social and cultural standpoint and are framed by assessment

7 For brevity, here "LVM" refers only to between-person LVMs.

designers and users in light of the students, the contexts and the purposes at issue. Task performances are interpreted in terms of choices, approaches and appropriateness as seen from that social standpoint.

The local, unique and multiply-determined socio-cognitive processes that produce learners' performances contrast starkly with the LVM formulation. If not as measurement, how are we to think of the model forms, the probabilities and the variables of an LVM in application? To the degree that a given LVM form and the variables adequately fit the observed $X$ values for collections of persons and tasks, a socio-cognitive interpretation is, as a data model, analogous to a mean–field approximation, which replaces many interactions with their average. That is, the fitted model provides probabilities for each observation in the *person–task ensemble* via the LVM form and estimated variables. The $\theta$s indicate data trends within the LVM form that are associated with persons and the $\beta$s indicate data trends associated with tasks. The probabilities given by $f$ are interpreted as the *modelers' descriptive probabilities for approximating observations in that person–task ensemble*, rather than as probabilities generated by hypothetical extant properties $\theta$ of persons and $\beta$ of tasks. These interpretations of model fit and variables depend on the socio-cultural milieu and personal histories of the individuals in the given ensemble (Byrne, 2002; Gong et al., 2023).

Hence, the contextuality of learning requires a re-conception of LVM symbol systems and their applications by regarding them as *descriptions of patterns in behaviour that emerge from multi-layered socio-cognitive processes*, which are embedded in complex linguistic and cultural contexts. This socio-cognitive perspective provides different narrative structures for organising and reasoning in educational assessment, even from the same learner performances, as they instantiate different arguments. This ontologically and epistemologically more elaborated understanding of LVMs, rather than their common (explicit or implicit) interpretation as reflecting personal properties, will lead to more appropriately—because contextually—grounded inferences in current practices in educational assessment (Mislevy, 2018).

The two previous contributions highlighted that careful, contextualised interpretations of psychometric results, such as using constructivist realist approaches, can enable meaningful applications of psychometric tests for pragmatic purposes in applied settings (e.g., legal, occupational). Psychological 'measurement', however, is widely used also in academic psychology to study individuals' behaviours, beliefs, abilities and other phenomena and to develop theories about them. Indeed, quantitative data generated with psychometric 'scales' form the basis of much of the empirical evidence used to test scientific hypotheses and theories in psychology. This requires critical analysis of the ways in which psychometric 'scales' and models are designed and which determine their appropriateness for empirical inquiry.

Specifically, let us set aside the ontological debate on whether psychical phenomena can have quantitative properties. Assuming they do, what properties must our approaches and methods have to be able to provide the epistemic evidence necessary to support this assumption? In other words, are the current theories and practices of psychological 'measurement' able to determine quantitative properties of psychical phenomena, if such exist, to warrant their interpretation as procedures of measurement?

## Statistics is not measurement: Psychologists confuse disparate epistemic activities thereby neglecting their actual study phenomena

Psychology's main approach to 'measurement' involves statistical, especially psychometric analyses, often likened to indirect measurement in physics given the non-observability of others' (e.g., participants') psychical phenomena. But statistics neither *is* measurement nor is statistics necessary for measurement. Physical measurement, even of non-observable properties (e.g., gravity on Earth), was successful long before statistics was developed (Abran et al., 2012; Chang, 2004).

In various transdisciplinary analyses, Jana Uher demonstrated that statistics and measurement involve disparate scientific activities for disparate epistemic purposes. Statistics deals with structural relations in data regardless of what these data represent. Measurement, by contrast, establishes traceable empirical relations between the specific quantities to be measured (the *measurands*) in the study phenomena (empirical or real study system) and the data and results (e.g., true scores) representing information about them (symbolic or formal study system). Hence, statistics concerns purely *syntactic* relations in a data set, whereas measurement also establishes the data's *empirical semantic* meaning regarding the real study phenomena to which these data refer and for which they (symbolically) stand (e.g., Uher, 2021a, 2022a,b, 2025).

## Psychometrics involves pragmatic result-dependent 'instrument' design and data modelling, which preclude realist inferences on the actual study phenomena

Psychometric 'instruments' (e.g., intelligence tests) are designed to discriminate well and consistently between cases (or groups) and in ways regarded important (e.g., social relevance). To achieve this, psychometricians align the structures of psychometric 'instruments', and those of the data that can be generated with them, to statistical criteria and operations (e.g., normal distributions, internal consistency, item discrimination). The assignment of numerical scores, as well, is aligned to the results' utility and pragmatic value. In intelligence tests, for example, IQ scores are assigned such as to inform about a person's deviation from the age-group specific average, which is set arbitrarily to 100 (and one standard deviation in the normal distribution is set arbitrarily to 15 in both directions). That is, these numerical assignments are aligned to practical purposes rather than to quantitative properties of the actual study phenomena. Indeed, given pronounced cohorts effects (e.g., age groups, Flynn effect; Flynn, 2012), persons with the same test performances may be assigned different IQ scores to enable comparisons with their specific cohort. That is, psychometric theories and empirical

practices are designed to generate results with pragmatic utility—they build on a *pragmatist framework* (Uher, 2021a, 2022b, 2023a, 2025).

*Pragmatism* is a philosophical perspective in which knowing the 'world' is understood as inseparable from human agency and practice within it. This often entails a focus on epistemology and methodology at the expense of ontology. This heterogeneous family of theories and beliefs involves a broad, historically shifting and in parts contrary range of interpretations, which is irrelevant here (Legg and Hookway, 2024). Yet some key features of pragmatism clearly apply to psychometrics. For example, the value of pragmatic research is judged by the effectiveness of its results for a specific problem (e.g., discriminating between individuals) rather than by the results' correspondence to some state of 'reality'. This contrasts with the various forms of realism, which emphasise the nature of 'reality' and specify our possibilities and limitations of generating knowledge about it (Mertens, 2023).

Psychometricians' pragmatic focus on the utility and practical consequences of empirical inquiry is evident in the targeted design of psychometric theories and practices to produce quantitative results that are useful for specific purposes (e.g., discriminating between cases). These *result-dependent pragmatic* approaches (Uher, 2021b) contrast with the widespread interpretation of psychometric results as reflecting structures in the actual study phenomena. 'Personality' or IQ scores, for example, are commonly interpreted as constituting results of 'measurement' and their quantitative information is attributed to the individuals under study (e.g., their 'psychophysical mechanisms' or intellectual abilities).

Such inferences, however, can be made *only* when systematic relations are established between the real study phenomena (empirical system) and the measurement results obtained about them in the formal (symbolic) system (Rosen, 1985, 1991; Uher, 2025). This presupposes the *realist framework* of measurement, which, however, is neither theoretically elaborated nor empirically implemented. Instead, psychometrics is centred on modelling data structures in the symbolic study system, whereas the relations between the real and the symbolic study system are being neglected (Uher, 2021a; see also Heine and Heene, 2025). Hence, there is a *gap* between psychometric results and the specific entities that are to be quantified in the actual study phenomena. Bridging this gap requires measurement.

## Measurement requires data generation processes that are traceable to empirical interactions with the study phenomena and to known quantity references

So, what is measurement? In her transdisciplinary analyses, Jana Uher highlighted that, despite fundamental differences in theories and practices, psychometricians' declared aims and result interpretations reflect basic ideas of measurement that are shared by metrologists, physicists and psychologists alike. These shared ideas can be formulated as two *basic criteria*, which distinguish, across the empirical sciences, measurement from other quantification practices that may be pragmatically useful but lack epistemic authority (e.g., evaluation). These epistemic criteria are (1) the justified attribution of results to the specific entities to be measured (*measurands*; e.g., an individual's duration of speaking in a situation) and (2) the public interpretability of the results' quantitative meaning regarding those measurands (e.g., *how* long that is). These criteria are not meant to classify approaches as 'superior' or 'inferior'. Rather, a *criterion-based approach to define measurement* is essential for scrutinising the epistemic fundamentals of a field's pertinent theories and practices, such as to highlight the epistemological inconsistencies inherent in psychometrics, and to pinpoint commonalities and inevitable differences between sciences (Uher, 2021c,d, 2023a).

To meet these epistemic criteria, empirical processes must build on two *corresponding methodological principles*, which underlie metrologists' frameworks of measurement and which are—on their abstract level of consideration—applicable across sciences. Accordingly, measurement requires documented, unbroken connection chains that establish proportional (quantitative) relations of the results with both the measurand's unknown quantity (e.g., in an individual; *principle of data generation traceability*) and a known quantity reference (e.g., international standard units; *principle of numerical traceability*; Uher, 2018a, 2020a, 2021b, 2022a,b, 2023a). These two types of traceability are established in iterative processes of theorising and empirical experimentation in which a real (empirical) and a formal (symbolic) study system, as well as their relations, are coherently related with one another. This *coordination* is crucial for justifying the assumption that a specific procedure does indeed allow us to measure a specific property in the absence of independent methods for measuring it as well as for justifying that specific quantity values are assigned to specific measurands. *Calibration* is used to refine the coordinated structure of a measurement process by specifying the ranges of uncertainties and errors for all its parameters to improve the accuracy of results (Chang, 2004; Luchetti, 2020; Tal, 2020).

Rosen's (1985, 1991) general model of measurement conceptualises this process as a *system*[8] *of four interrelated modelling relations*, comprising the (1) objects of research, (2) data generation (encoding), (3) formal manipulation (e.g., statistical analysis) and (4) result interpretation regarding the objects studied (decoding; Figure 2). This involves modelling the presumed relations within the real study system, comprising the non-observable object of research (measurand), the object used as instrument (including a known reference quantity) and the observable indication produced from their (non-observable) empirical interaction. Their presumed causal relations (arrow 1) are then explored empirically through unbroken and traceable relations to, within and back from the formal system that is used to study that real system (arrows 2, 3 and 4). In iterative feedback loops, the four modelling relations in Rosen's system (arrows 1 to 4) are passed through over and over again, thereby *re-coordinating* and *re-calibrating* them with one another until they

8 Rosen (1985; 1999) himself and others refer to this process model solely as *modelling relation*. To highlight that it involves the coherent modelling of four interrelated modelling relations (arrows 1 to 4) and to pinpoint key distinctions to the statistical modelling of data, which concerns solely arrow 3 in Rosen's general model, Uher (2025) refers to his process model as a *system of interrelated modelling relations*.

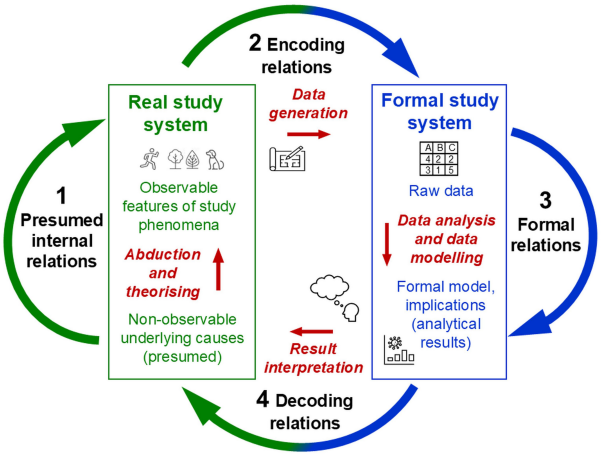



FIGURE 2
Rosen's general process structure of measurement involves a coherent system of four interrelated modelling relations. It conceptualises the real (empirical) study system and the formal (symbolic) system used for studying it as mathematical objects as well as the processes (mappings, relations) each within and back and forth between them, depicted as arrows. © From Uher (2025), Figure 1; adapted from Rosen (1985).

are theoretically and empirically coherent, indicating successful modelling of the real study system (for details, see Uher, 2025). Coordinated and calibrated processes enable epistemically justified attributions of the results to the quantities to be measured in the study phenomena (criterion 1) as well as the public interpretability of the results' quantitative meaning regarding those measurands (criterion 2).

Rosen's general process scheme shows that, by focusing on statistical modelling (arrow 3, Figure 2), psychometricians neglect the three other modelling relations (arrows 1, 2 and 4) without which a formal model cannot be coordinated and calibrated with the real study system. Their interrelations are neither conceptualised nor empirically established through traceable connections but simply decreed in psychometricians' result interpretations, declared aims and operationalist procedures of numerical assignments (Uher, 2025).

### Pragmatic quantifications with predictive power but without explanation

Quantitative psychologists' 'measurement' jargon alludes to the *epistemic authority* of genuine measurement yet without fulfilling the necessary criteria. This misleads the public, practitioners and scientists because, in both everyday life and science, the term measurement implies that some part of 'reality' (e.g., a bottle's volume) is being quantified in justified and verifiable ways. Therefore, we trust measurement results (e.g., volume indications on wine bottles; criterion 1) and can interpret (with the relevant knowledge) the specific quantitative meaning that they have for the object measured (e.g., how much '75cl' is; criterion 2).

Approaches of psychological 'measurement' (e.g., psychometrics), by contrast, allow for generating *pragmatic quantifications* that are useful for distinguishing individuals by their observable responses or performances and for making decisions and predictions on the basis of the differences and relations observed. But these approaches do not constitute measurement because they fail to establish coherent relations to the study phenomena both theoretically and empirically. By adapting the 'instruments' and results instead to statistically useful data structures, these result-dependent approaches cannot explore the observed responses or performances for their underlying causes, such as what specific intellectual abilities individuals may use to show a specific performance in a given task.

The lack of epistemic validity also compromises psychology's efforts to tackle its crises (e.g., replicability). Current initiatives (e.g., robust statistics, replication) solely concern practices focussed on data analysis and interpretation. But psychology's crises cannot be solved without transparency in its data generation (Uher, 2023a). Without advancing *ontological concepts and theories* about the study phenomena (e.g., individuals' thought processes, constructs, behaviours; Uher, 2013, 2015a,d, 2016a,b, 2021c, 2023b) and without elaborating *epistemological and methodological concepts* of how relevant features of these phenomena can be made amenable to quantitative investigation, and if at all (Uher, 2015b,c, 2018a, 2019, 2021d), the root causes of replicable quantitative findings cannot be identified (see Topic 1).

Psychology must tackle the *gap* that often exists between its quantitative findings and statistical models, on the one side, and its actual study phenomena and the specific quantities to be measured in them (measurands), on the other. Therefore, *genuine analogues* of measurement must be advanced for which Rosen's process scheme of measurement and the transdisciplinary concepts of data generation traceability, numerical traceability and the two epistemic criteria of measurement are useful. Clinical research (e.g., on quality of life, chronic disease, therapeutic efficacy) has already pioneered successful implementation of such approaches and advanced their epistemic fundamentals (for details, see Uher, 2025).

This epistemic gap is often overlooked, however, because many psychologists mistake the *inbuilt semantics* of their language-based methods—thus, descriptions of their study phenomena (e.g., in rating scales, item variables, statistical models)—for the phenomena described (Uher, 2025). This shifts our focus to psychology's means of scientific inquiry and their distinction from the study phenomena, as we discuss now in Topic 3.

## Topic 3: Peculiarities of psychology's study phenomena and its means of scientific inquiry: Constructs and language-based methods

Psychology's study phenomena feature peculiarities, such as emergence of novel properties that feed back to and change the very processes from which they emerge in multi-level feedback-loops, leading to continuous change and development and thus, to higher-order complexity (see Topic 1). Such peculiarities are not known from the non-living 'world' studied in physics and metrology. Moreover, psychology explores not

just objects and relations of specific phenomena in themselves (e.g., behaviours) but also, and in particular, their *individual (subjective) and socio-cultural (inter-subjective)* perception, interpretation, apprehension and appraisal (Wundt, 1897; Uher, 2021c, 2025). These complex study phenomena are described in *constructs*.

> "A *construct* is a conceptual system that *refers* to a set of entities—the *construct referents*—that are regarded as meaningfully related in some ways or for some purpose although they actually *never occur all at once* and that are therefore considered only on more abstract levels as a joint entity (italics as in original; Uher, 2022b, p. 14).

All humans develop and intuitively use constructs in everyday life (Kelly, 1955, 1963). Everyday psychology is replete with constructs, which are encoded in everyday language (Vygotsky, 1962). That is, constructs form an important part of our human thinking. Constructs are also important conceptual means of scientific inquiry in psychology (e.g., 'intelligence', 'leadership', 'benevolence') and the social sciences (e.g., 'power', 'democracy'). Each construct refers to a theoretical universe of referents that are jointly considered for a purpose (e.g., evaluation, explanation) and from a specific viewpoint (e.g., normativity, specific theory) but that can never be observed all at once—constructs are *multi-referential* conceptual systems. For empirical studies, a manageable subset of referents is chosen to serve as *indicators* (Uher, 2022b, 2023b). To conceptually handle constructs, given their level of abstraction, language plays a crucial role in their description and empirical investigation. The distinction between constructs and their referents (e.g., empirical indicators) as well as the intricacies of human language, however, involve complexities that present unparalleled challenges to quantitative inquiry.

## Psychologists' cardinal error: Confusing ontological with epistemological concepts

In her transdisciplinary research, Jana Uher highlighted that, ontologically, all phenomena can be described in their being. To elaborate how knowledge about a given study phenomenon can be gained, thus epistemologically, scientists must decide, in every study, which specific phenomena they aim to explore and which ones they use as epistemic means for exploring these study phenomena. The necessity of this epistemic distinction, first recognised in quantum physics (Heisenberg, 1927), is not well considered in psychology (Uher, 2025). Moreover, this distinction is particularly intricate in psychology given the *anthropogenicity of science*—the fact that all science is made by and for humans using the abilities of the human mind (e.g., conceptualising, generalising, abstracting; Uher, 2022b, 2023a,b). Empirical science is experience-based by definition (from Greek *empeiria* for experience). For scientists exploring mind and experience, this complicates the logical distinction between the specific psych*ical* (e.g., mental) phenomena that they aim to study as their objects of research (e.g., participants' beliefs, abilities, folk constructs) and those psychical (and further) phenomena that they use as epistemic means to investigate the study phenomena (e.g., psychologists' own inferences, theories, methods, Big Five constructs). These epistemic means of inquiry are properly termed psych*ological*[9], derived from Greek *-logia* for body and theory of knowledge (Lewin, 1936; Uher, 2021b, 2023a).

Failure to make the crucial epistemic distinction between the study phenomena and the study means (in a study) entails the confusion of ontological with epistemological concepts—therefore, it is termed *psychologists' cardinal error* (Uher, 2022b). This logical error makes the distinction of disparate scientific activities (e.g., theoretical vs. operational construct definition) technically impossible, thereby distorting scientific concepts and procedures (Uher, 2013, 2015a,b,c, 2023b). This error can occur in various parts of the empirical research process.

### Conflations of the study phenomena with the study means masked and perpetuated by psychological jargon

Psychologists' cardinal error occurs when psychologists use key terms ambiguously (e.g., 'constructs', 'variables', 'attributes'), thereby *conflating the study phenomena with study means*. Constructs, for example, are often mistaken for the study phenomena to which they refer (*construct–referent conflation*; Lovasz and Slaney, 2013; Maraun and Halpin, 2008; Maraun and Gabriel, 2013; Slaney, 2017; Uher, 2013, 2021a,b). This leads many to confuse the abstract concept of 'intelligence' with the various intellectual abilities to which it refers and that never occur all at once but that are just jointly considered for some purpose. This logical error is promoted by the operationalist idea that a study phenomenon's theoretical meaning could be established through the empirical operations that are used to investigate, manipulate or elicit it. Specifying operational procedures may help to pilot conceptual research. But ultimately, operational specifications must be replaced by proper theoretical definitions of the study phenomenon (Green, 2001; Feest, 2005). If these distinctions are not made, further logical errors occur. For example, when reasoning ability is operationally 'defined' as test performance, this ability cannot also be used to explain this performance. A phenomenon cannot be defined by its effects. Such assumptions *conflate cause with effect*, thereby *turning the effect into its cause* (Hibberd, 2019; Uher, 2022b).

These logical errors also occur when—misled by the availability of single word terms (e.g., 'personality')—researchers treat constructs as real entities, thereby turning abstract ideas into things (entification, reification, hypostatic abstraction; Peirce, 1958, CP 4.227). This occurs, for example, when 'personality' constructs are interpreted as entities residing in individuals (e.g., 'psychophysical mechanisms') that causally underlie their behaviours, feelings and thinking—thereby *turning the description of study phenomena into their explanation* (Uher, 2013). Further logical errors arise from intricacies of human languages that are not well considered in psychology.

9 Attentive readers will have noticed that the psychical—psychological (psyche—psychology) distinction is not made consistently in this paper, which reflects the differences in our linguistic and conceptual habits as authors (also given different language backgrounds).

## The intricacies of human languages

Language is humanity's greatest invention (Deutscher, 2006). With words, we can refer to objects of consideration even in their absence (meaning), although what we say or write (signifiers) typically bears no inherent relations (e.g., resemblance) to the objects referred (referents). This representational function of language is built into its *semantics*—the rules that specify the meanings that words, phrases and sentences conventionally convey in terms of what they refer to and stand for in the real 'world' (their referents). The complex rules of languages (e.g., semantics, syntax, pragmatics)—developed in socio-linguistic communities and internalised during language socialisation—mediate and shape intra-individual and inter-individual processes (e.g., thinking, interacting). Therefore, language and psyche are inseparable from one another, while still constituting different kinds of phenomena (Peirce, 1958; Uher, 2015a,b, 2016a, 2018a; Valsiner, 2000, 2007; Vygotsky, 1962). Because of this entanglement, we do not perceive our words just as tokens of the objects to which they refer but as these objects themselves. Therefore, in our minds, we easily mistake the word for the thing, the map for the territory, the menu for the food—the 'world' as it is with the 'world' as it is thought about and described (Uher, 2025).

Our human tendency to mistake verbal descriptions for the phenomena described leads to further instances of psychologists' cardinal error. These occur when researchers—distracted by the ease of using language and unaware of its inherently representational nature—focus only on the *inbuilt semantics* of language, thus on the meanings that words and statements generally have (Uher, 2025). This often obscures the epistemic necessity to distinguish the study phenomena (e.g., individuals' feelings) from their verbal description in the language-based methods used for studying these phenomena (e.g., item 'scales', variable names), leading to the confusion of ontological and epistemological concepts. This cardinal error often underlies evaluations of face validity and content validity of psychometric 'instruments'. It also underlies the widespread *nominalism* in quantitative psychology—the belief that any method that is *nominally (by name)* associated with a study phenomenon could be epistemically valid for empirically studying it (e.g., 'anxiety scale', 'openness scale'). This contributes to the proliferation of overlapping 'scales' (e.g., various 'anxiety scales') and of the likewise overlapping constructs that their items are meant to operationally define (Sechrest et al., 1996; Toomela, 2010; Uher, 2021b, 2022b).

The *inbuilt semantics* of language also often leads psychologists to misinterpret raters' judgements of verbal statements as measurements of the phenomena described in those statements. The epistemic necessity to establish traceable coordinated and calibrated relations between the symbolic and the empirical study system gets out of focus (Uher, 2025). This entails the risk of replicating just verbal descriptions instead of exploring the real phenomena for which these are meant to stand. Therefore, quantitative psychology is at risk of doing *pseudo-empirical* research, which mostly re-discovers what is necessarily true given the logico-semantic relations built into its language-based methods (Arnulf et al., 2024; Shweder, 1977; Shweder and D'Andrade, 1980; Smedslund et al., 2022; Smedslund, 1991, 2016). Indeed, many overlook that human languages have socio-culturally constructed structures and meanings, which do not derive from the ontic 'reality' that they describe and which therefore vary considerably between languages (Deutscher, 2010; Boroditsky, 2018; Uher, 2025).

This also entails challenges also for philosophy of science. For example, some realist perspectives explicitly involve the presupposition that 'reality' is "mind-independent" and "language-independent". These terms, however, if taken literally, may create the illusion that minds and languages could be generally independent of and thus, extraneous to 'reality' rather than forming part of it as well. This is particularly misleading for psychologists who aim to explore the 'reality' of mind and whose primary means of empirical inquiry is language, which, moreover, is internalised in human minds. Instead, it is crucial to specify, which parts of 'reality' are meant to be studied and which parts of 'reality' are used as epistemic means for exploring these study phenomena—thus, to distinguish ontological from epistemological concepts (e.g., psych*ical* from psych*ological*; Uher, 2023a).

To scrutinise the *epistemic role of language* in empirical inquiry, it is important to ontologically study its elements, structures and relations. Linguists, information scientists, artificial intelligence researchers and other scholars established *ontologies of language* that describe its syntax and inbuilt semantics (e.g., using digital networks), such as those underlying natural language processing (NLP) systems and large language models (LLMs). Our next contribution demonstrates how language ontologies can elucidate some key problems in quantitative psychology and highlights fundamental issues still hardly considered.

# The semantic representations of psychological phenomena reappear in statistical data as self-reinforcing ontologies

All scientific psychological phenomena have in common that they also exist as linguistically defined topics of research. Most psychological constructs also appear as topics in everyday conversation and public discourse. The relationship between psychologically theorised and linguistically defined 'constructs', on the one hand, and their purported ontological 'reality', on the other, remains elusive. It has regained importance, however, through the development of digital language processing techniques, as Jan Ketil Arnulf and colleagues documented in their line of research around the Semantic Theory of Survey Response (Arnulf et al., 2014, 2018).

## Constructs as representations in language models

While early 20th century psychology displayed a sound scepticism towards 'mentalistic' concepts as legitimate objects for scientific scrutiny, the behaviourist reaction equally created overly strict criteria for legitimate research topics. In the 1950s, the American Psychological Association (APA) accepted in its methods standards the adoption of 'latent constructs' to the extent that these could be legitimised by statistical modelling techniques (Slaney, 2017). Since then, the domain of psychology has expanded with

a growing range of non-observable phenomena that mainly exist through their statistical properties in empirically collected data (Larsen et al., 2013; Lamiell, 2013, 2019a; Smedslund, 2021).

However, theoretical doubts about the ontological status of such constructs and their purported relationships have repeatedly been raised. Most importantly, it has been shown that their empirical relationships, in many cases, may be not empirical but pre-given through their logical or semantic relationships—and thus, pseudo-empirical and tautological (Semin, 1989; Smedslund, 1991, 2012, 2016).

These concerns have rarely been addressed so far. Instead, ever-increasing statistical sophistication and primarily language-based methods (e.g., rating 'scales') have been used to establish ever more 'latent constructs' in psychology. This has continued without ascertaining the nature of the phenomena and processes involved in generating the data that serve as input to the statistical models. With the emergence of natural language processing algorithms and software, this concern has now been turned into an empirical investigation. It is possible to use verbal 'measurement scales', variables and construct definitions as well as other methodological features as input to text algorithmic analysis (Arnulf et al., 2021). These technologies were originally built on Latent Semantic Analysis (LSA) but have later become much more precise through the adoption of more advanced language models, such as BERT (Bidirectional Encoder Representations from Transformers).

The key point of this approach is that psychology has overlooked how language itself is describable as having mathematical features. The mathematical features of meaning in language are precisely what enable the powerful large language models (LLMs) that are now ubiquitously available (Devlin et al., 2018; Landauer and Dumais, 1997)—often referred to as "Generative Artificial Intelligence" (genAI; Chang et al., 2024). The semantic approach to the measurement problem in psychology is that the sampled statistics will easily reflect *what we say about* a phenomenon—rather than the phenomenon itself—unless special attention is taken to avoid it (see Topic 2).

The empirical proof of this claim is built on the fact that digital text analysis allows the replication of statistical psychometric models using only textual data as inputs and without any involvement of research participants using these verbal 'scales' to make quantitative assessments—thus, without any empirical investigation. It is possible to show that much of the systematic information captured by psychometric modelling stems from the semantic patterns of construct definitions and verbal 'measurement scales' as well as from their mutual relationships (Arnulf and Larsen, 2021; Arnulf et al., 2024).

## Statistical features of constructs do not make them true or false

Semantically derived findings have two problematic implications for science: first, they are predictable a-priori (Wierzbicka, 1996; Smedslund, 1978, 2016) and therefore do not expand our knowledge. Second, their empirical status remains untested because it is possible to make both true and false statements in language. One such implication occurs in cross-cultural studies on leadership where it was found that propositions about leadership correlated in the same way across the 'world', even if local behaviours by people in workplaces might be very different (Arnulf and Larsen, 2020).

From a measurement perspective, it can be shown that the quantitative information (data) commonly used to legitimise the ontological status of many 'latent' psychological constructs does not stem from some unobservable psychological study phenomena. Instead, the quantitative relationships are features of the linguistic structures that we use to represent these study phenomena in operationalisations and variables (Arnulf et al., 2018). When this happens, psychometric models reflect the ways in which researchers and participants describe human experience, emotions, thinking and other psychological phenomena. Ascribing these statistical properties to independently existing phenomena extraneous to language is an error of category, mistaking the representations for the represented—the menu for the food (Arnulf et al., 2024).

## The human struggle to discern empirical from semantic problems

What makes this error practically possible may be the social construction of human 'reality', turning many constructs into realities by simply treating them as real entities (reification, entification). This obstructs our view of many such constructs as historically developed, belonging to socio-cultural, professional or other communities of practice. However, it can be shown that this semantic nature of the subject matter effectively locks psychological research in mutually defining semantic networks, which can be visualised in graphical networks (for an example, see Figure 3). The conventions of factor analysis restrict the explained variance of its results to an average of 42%, above which explanations appear as auto-correlations and as uninteresting if they become much lower (Smedslund et al., 2022). Since the 1950s, the combination of construct validation conventions and semantic networks has turned psychological research into a self-perpetuating machine that keeps explaining semantic phenomena by rephrasing them as other constructs or other operationalisations instead of tapping into their underlying realities—a mistake of categories.

Within this natural language processing (NLP) paradigm, now enabled through powerful algorithms and software systems, one of the most pressing psychological research questions is to explore why humans in general—and researchers in particular—lose sight of the semantically given frameworks of our socio-linguistically constructed 'world' so easily. As psychologists, we must better understand why we struggle to differentiate empirical from semantic research problems. This opens up novel perspectives on psychology's crises in replicability, validity and generalisability as well as on the role that psychologists themselves may play in their perpetuation.

All words have meaning. The meaning of every word is a construct (Vygotsky, 1962). Exploring the role of constructs in psychological research requires an elaborated *ontology of constructs* (Kelly, 1955, 1963; Uher, 2023b). Constructs are also studied outside of psychology, such as in information science as in our next contribution, which analyses constructs using Mario Bunge's

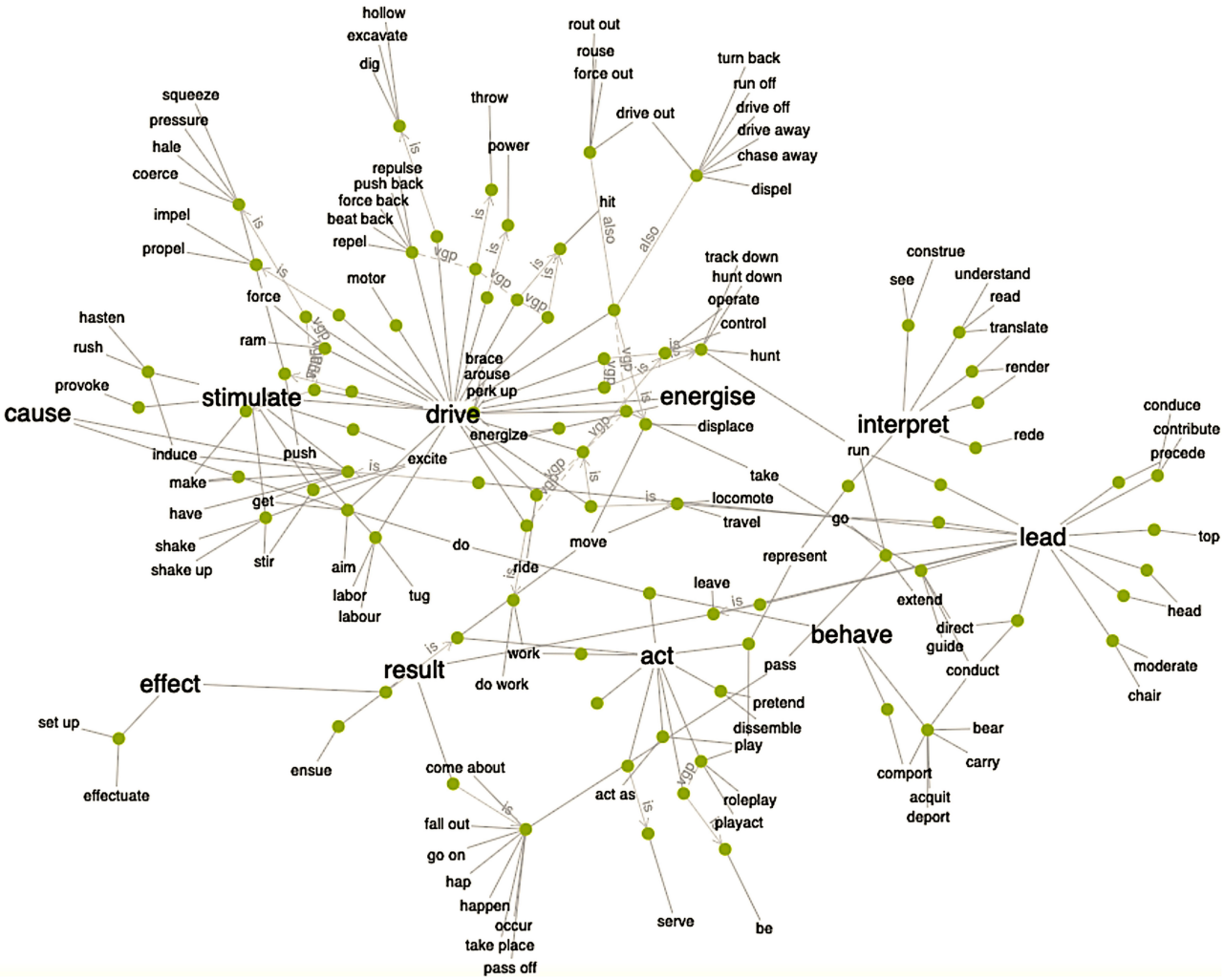



FIGURE 3
Semantic networks Natural language terms mutually define each other within chains of semantic relationships. When used in language-based methods, such as in descriptions of the study phenomena in rating items, semantic relationships were shown to reappear as 'explained variance' in statistical research models.

philosophy. Bunge advocated for scientific realism, positing the existence of a "mind-independent[10]" 'reality' that can be known and described, at least up to a point (Bunge, 1977, 1993). Through experience, reason, imagination and criticism, we can obtain some truthful knowledge about this 'reality', which, although variously problematic (e.g., abstract, incomplete, fallible), can also be improved (Bunge, 1993; Cordero, 2012; Mahner, 2021). Bunge elaborated a *materialist ontology*, founded on the presupposition that the real 'world' (what exists) is composed only of material things. Things can change (construed as events) and possess properties that characterise them. Interactions between things form systems that have novel emergent properties. The real 'world' is therefore a 'world' of systems. Bunge conceptualised the 'mind' not as a thing but as mental properties of complex brains, which emerge from processes of neuronal systems. This *emergentist materialism* thus rejects a dualist body–mind ontology (Bunge, 1981; Mahner, 2021). Our following contribution applies Bunge's ontology to elaborate on constructs and their relations to their indicators as well as to the 'instruments' that are used for empirical explorations.

## An ontological analysis of construct–indicator and indicator–instrument relationships: Novel theoretical perspectives on current controversies

Constructs and their indicators are central to theory building and theory testing in many disciplines. *Theories* articulate relationships among constructs. *Indicators* are used to measure

10 See our previous discussions on the notion of a "mind-independent reality", which, if taken literally, can be misread as (e.g., the researcher's) mind being generally independent of, thus, extraneous to 'reality' rather than forming part of it as well (p. 23).

construct values. Yet the nature of constructs and the relationships among them as well as the nature of indicators and the relationships between constructs and indicators remain contested. The controversies that have occurred are unlikely to abate until the ontological assumptions that underpin constructs and indicators are surfaced and scrutinised (Bagozzi, 2011; Borsboom, 2005; MacKenzie et al., 2011). In this light, Ron Weber used Bunge's (1977, 1979) materialist ontology to analyse the essential nature of constructs and indicators (Weber, 2012, 2021). He chose Bunge's ontology because it is comprehensive, formalised and widely used (Matthews, 2019).

## Ontological fundamentals: Objects, things, constructs and properties

The fundamental unit in Bunge's ontology is an *object* defined as "whatever can exist, be thought about, talked about, or acted upon. The most basic, abstract, and general of all philosophical concepts, hence undefinable. … Objects can be individuals or collections, concrete (material) or abstract (ideal), natural or artificial" (Bunge, 2003, p. 199). He divides objects into two ontological categories: things and constructs. *Things* are objects in the 'world' that exist independently of their perception and conception by sentient beings (which are things themselves as well). *Constructs* are objects that exist in sentient beings' brains. As sentient beings, we cannot perceive the 'world' directly; we perceive it only though our constructs. Hence, whenever we talk about things, we actually talk about our *models* of things—the constructs that we use to comprehend the 'world'.

The traits that characterise a thing or construct are its *properties* (Bunge, 1977). Two types of properties exist in relation to things. *Properties in general* are common to a class of things. For instance, scholars might study a general property called "benevolence" and the extent to which it is possessed by a class of humans called "managers" (Serva et al., 2005). *Properties in particular* are the specific levels (values) that specific things in a class possess of a given general property. For example, the specific level of benevolence (e.g., "high") possessed by a specific manager called "Jane" is the particular property of a specific thing from the class called "managers". Weber (2012) argued, however, that, during theory building and testing, scholars often unwittingly tend to use the term "construct" in a more specific way than Bunge and use it to mean a *property in general* of a thing.

## The ontological nature of indicators and their relationship to constructs

During theory testing, some focal constructs (properties in general) can be measured directly (e.g., a person's height with a ruler). Often, however, focal constructs are unobservable and must be measured indirectly. Indirect measurements of constructs occur via indicators, which are sometimes observable proxies for the unobservable focal construct (e.g., weight as an indicator of a person's stress level; Bunge, 2010). In psychology and the social sciences, however, indicators are often unobservable in themselves as well. Therefore, they must also be measured indirectly (e.g., managers' typical ways of acting over some time). Nonetheless, scholars might deem that using a set of indicators that can be measured only indirectly (and combining them in some way to determine the focal construct's value) provides the best measure of that focal construct.

Using Bunge's ontology, Weber (2021) argued that scholars predominantly conceive indicators, often unwittingly, as general properties of some class of things. For instance, scholars might study the focal construct "benevolence" as a general property of a class called managers, and they might choose another set of general properties as indicators of that construct to obtain an indirect measurement of it. Indicators of the focal construct "benevolence" might be managerial actions, such as looking out for important issues, ascribing importance to needs and desires and going out of the way to help (Serva et al., 2005). The specific level of "benevolence" for Jane as a specific manager (particular property) will be determined on the basis of her specific levels measured for each of these three indicators.

## The ontological nature of instruments and their relationships to construct indicators in measurement

To measure the values of indicators for specific things, such as for specific persons (i.e., particular properties of a particular person), scholars use *instruments*. Under Bunge's ontology, instruments are also things with properties. For instance, a questionnaire[11] (a thing) for studying the focal construct "benevolence" (property in general) of managers (things) might have several manager-descriptive indicators comprising item statements with Likert rating 'scales'. The item statements themselves (without any specific Likert 'scale' rating) are properties in general of the questionnaire instrument (thing). Observers (e.g., a manager's subordinates) make judgements about the levels of these indicators (properties in particular) on the basis of their perceptions of their manager's actions. Three such indicators might be "looks out for important issues", "ascribes importance to needs and desires" and "went out of the way to help". Subordinates use these indicators with the Likert 'scales' to rate their perceptions of their manager's actions. The indicators with specific Likert 'scale' ratings (e.g., "3", "6") are the questionnaire's properties in particular.

Ideally, the values that an indicator (or set of indicators combined) assumes for specific things should be *isomorphic* with the values that the focal construct assumes for these things (Borsboom, 2008). In this regard, ideally, an *auxiliary theory* should have been developed to explain why specific indicator values obtained via a measurement instrument are isomorphic with the focal construct's values (Bunge, 1974, 1975, 2010; Edwards, 2011).

11 Questionnaires and Likert 'scales' are used here to illustrate key concepts because psychologists are familiar with them. The preceding sections have already highlighted these methods' serious limitations as 'measuring instruments', given their measurement theoretical, conceptual and methodological deficiencies as well as the intricacies that natural language entails for language-based methods (see Topics 2 and 3).

## Property scopes, property pre-orders and measurement instruments

Scholars strive to design and use high-quality instruments to measure the particular properties of specific things (e.g., specific behavioural actions of specific persons)—the measurands. Therefore, many method researchers focus on developing instruments that produce 'valid' and 'reliable' measures of focal constructs (Straub et al., 2004). Weber (2021) argued, however, that this literature is fraught with ambiguities and inconsistencies. Moreover, some approaches to measurement are highly contested—for example, whether formative instead of reflective indicators[12] should ever be used to measure the value of constructs (Bollen and Diamantopoulos, 2017; Guyon, 2018; Hardin and Marcoulides, 2011).

Weber (2021) proposed a new way to conceive and choose indicators on the basis of Bunge's ontology and Bunge's notion of the *scope of a property*, which is the set of all real-world things that possess that property. For instance, the scope of the property "benevolence" is the set of all individuals who possess it (at some level). If the scope of a property is a single thing, however, the property is possessed only by that thing (it is unique to that thing). Because different properties have different scopes, they apply to different subclasses of things. In a given class of things, these scopes therefore enact a *pre-order* (reflexive and transitive) on the given properties (Bunge, 1977). For example, in a putative theory about "manager trustworthiness", the scope of the property "benevolence" might be hypothesised to be a *subset* of the scope of the property "helpful" (Serva et al., 2005). That is, *some* but not all managers who go out of their way to help others are also "benevolent" (necessary condition), whereas *all* managers who are "benevolent" also go out of their way to help (sufficient condition). In Bungean terms, the property "helpful" *precedes* the property "benevolence" and the property "benevolence" *succeeds* the property "helpful". Property scopes and the property pre-orders that they entail can be visualised in Venn diagrams (Figure 4).

Importantly, Weber (2021) showed how the notion of property scope motivates new ways to assess the quality of a set of indicators, such as their *scope validity*. Specifically, if the set of indicators used to measure a focal construct *precede* that construct, ideally the *intersection* of the scopes of these indicators will equal that focal construct's scope. Alternatively, if the set of indicators used to measure a focal construct *succeed* that construct, ideally, the *union* of the indicators' scopes will equal that construct's scope. Weber highlighted that the importance of scope validity is primary to the importance of traditional instrument validity and reliability measures. That is, if an instrument does not have scope validity in the first place, its use can lead to "false positive" and "false negative" outcomes, although the instrument might have high levels of convergent and discriminant validity.

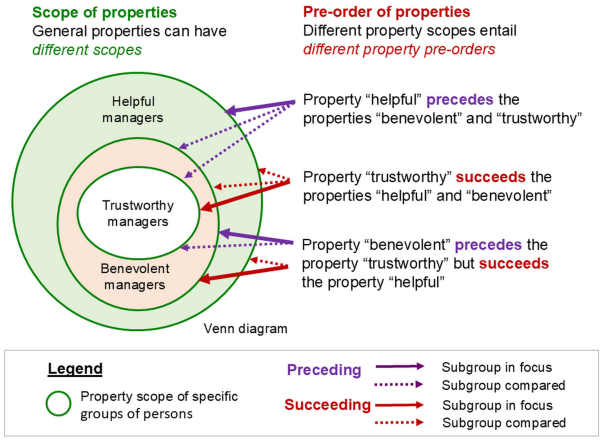



FIGURE 4
Property scopes and property pre-orders associated with subclasses of things (managers). The different scopes of properties enact different pre-orders on these properties. The property "trustworthy" applies to a subclass of all managers who are "benevolent" of all those managers who are "helpful". That is, the property "benevolent" precedes the property "trustworthy" and succeeds the property "helpful". Adapted from Weber (2021).

## Choosing indicators on the basis of property scopes and property pre-orders

When designing or choosing an instrument, scholars must evaluate carefully whether the indicators precede or succeed the focal construct in the pre-order of the properties included by that focal construct. They must then try to determine these indicators' likely scope. If scholars conclude that the intersection of the scopes of preceding properties and the union of the scopes of succeeding properties do not equal the scope of the focal construct, the designed or chosen instrument might not yield valid measures of that focal construct (Weber, 2021).

These ontological concepts from information science can provide novel perspectives also for one of quantitative psychology's most pervasive problems—the approaches for generalising findings across individuals that we discuss now in our next Topic 4.

# Topic 4: Psychology's approaches for generalising findings across unique individuals: Common errors and epistemically justified alternatives

The question of how we can develop general knowledge and universal categories given that we can always observe only particulars—the problem of universals (see Topic 2)—is of specific relevance for psychology as a science studying unique individuals. Quantitative psychologists, especially those building (implicitly) on positivist approaches (see Topic 2), commonly use statistical sample-level findings to generalise across individuals. Epidemiologists and health scientists, by contrast, are long wary of different types of fallacies that inferences from groups (on different levels of aggregation) to single cases, and vice versa, may

12 *Formative indicators* are hypothesised to causally affect the latent construct that they underpin, whereas *reflective indicators* are hypothesised to be affected by the latent construct that underpins them. That is, the direction of causality between constructs and indicators differs between formative and reflective indicators (Bollen and Diamantopoulos, 2017).

entail (Diez Roux, 2002). Quantitative psychologists, however, seem still oblivious of the problematic fundamentals on which the use of sample-level statistics for studying individual-level phenomena are based. Therefore, let us first scrutinise the underlying methodological, epistemological and ontological presumptions.

## The overlooked non-ergodicity of psychology's study phenomena: Why sample-level statistics cannot enable individual-level explorations

The advent of the assessment industry (e.g., in the American military in WWI; Gould, 1996), group-based experiments (Danziger, 1985a), rating methods (Thurstone, 1928; Likert, 1932) and statistical advances (Michell, 2023; Spearman, 1904) shifted psychologists' original focus on analysing psychical processes in individuals—psychology's *theoretical unit of analysis*—to analysing distribution patterns in populations, which became psychologists' primary *empirical unit of analysis*. Now, results were presented as aggregate data obtained from many individuals (e.g., group averages) yet without analysing individual patterns (Danziger, 1985b; Lamiell, 2019b). Still, psychologists continued to interpret their findings with regard to single individuals, which remained their focus of interest and *theoretical unit of analysis*. Personality psychologists, for example, commonly equate between-individual differences with individuality ('personality') and use sample-level statistics (e.g., factor analysis) to 'study' intra-individual functioning and development (e.g., using the Five Factor Model of 'personality'; Lamiell, 2013; Uher, 2018c, 2022b).

Inferences from sample-level findings to individual-level phenomena presuppose *ergodicity*—a property of stochastic processes and dynamic systems, which involves that their elements' synchronic and diachronic variations are statistically isomorphic. Ergodicity fits all invariant phenomena, which do not change and develop and in which simultaneity and successivity are therefore statistically equal (e.g., in some inanimate systems). Human individuals, however, are not all the same. Individuals, and the phenomena studied in them (e.g., behaviour, experience, language), vary, change and develop—thus, they change momentarily and over periods of time both intra-individually and inter-individually. Almost a century ago, the mathematicians Birkhoff (1931), John von Neumann and others advanced *ergodic theory*, a branch of mathematics originating in statistical physics (Gray, 1988). Using classical mathematical-statistical (ergodic) theorems, they proved that sample-level findings (e.g., group comparisons or correlations) can be generalised to single cases (e.g., individuals) *only if* (1) each case obeys the same statistical model (*homogeneity assumption*), and (2) the statistical properties (e.g., factor loadings) are the same at all points in time (*stationarity assumption;* Molenaar and Campbell, 2009). Why did ergodic theory elude quantitative psychologists, despite their keen interest in implementing mathematical-statistical approaches analogous to the physical sciences (Uher, 2022b)?

Presumptions of ergodicity are logically necessary for sample-to-individual inferences as well as pragmatically and methodically convenient. But they are invalidated already by ordinary everyday experience—not to mention an established body of empirical and theoretical research in psychology (e.g., Molenaar, 2004, 2008; Molenaar and Campbell, 2009; Richters, 2021; Salvatore and Valsiner, 2010; Speelman and McGann, 2020; Valsiner, 2014b; van Geert, 2011). The assumption of psychical homogeneity also contradicts fundamental design principles underlying all complex living systems in which different (non-isomorphic) structural elements are capable of performing or contributing to the same function, and vice versa, the same structures to different functions. That is, complex living systems feature both many-to-one structure–function relations (*degeneracy*, e.g., polygenic 'traits') and one-to-many structure–function relations (*pluripotency*, e.g., pleiotropic 'genes'; Mason, 2010, 2015). These unifying explanatory principles underlie the psychological concepts of *equifinality and multifinality*—individuals' capacities to leverage different psychical processes and structures to accomplish the same behavioural outcome, and vice versa (see Topic 1; Richters, 2021; Sato et al., 2009; Toomela, 2008b; Uher, 2022b, 2025).

When psychologists ignore their study phenomena's non-ergodicity in their statistical analysis, this entails fallible inferences as our next contribution shows. It highlights their implications for the interpretation of psychological findings and their replicability and presents an analytical method that allows for mitigating them.

## The ergodic fallacy: How psychology's erroneous ergodic assumptions can explain its inferential and reproducibility issues

Typical practice in psychological research is to aggregate data from many individuals to enable statistical analysis and to draw conclusions. In particular, the averages of scores of performances, or other psychological variables, are used to make inferences about the group of individuals studied—and even about the entire population from which it was sampled. These inferences are typically made in the form of generic statements about how "people" generally behave. These inferences are then used to make predictions about what single individuals might do in certain circumstances. Craig Speelman and Marek McGann articulated many problems with this chain of inferences, building on longstanding work across psychology's history.

### Implicit assumptions of ergodicity entail fallible inferences from empirical findings, obscured by generically worded conclusions

Speelman and McGann (2013) highlighted several assumptions underlying the use of averages, which are often implicit and almost always problematic. Most vital is the idea that averaging removes noise in a data set to provide a 'clearer' picture of some 'true' value. Variance around the mean is supposed to originate from unimportant or possibly random factors that can be 'averaged out' by focussing on the central tendency. This builds on the implicit assumption that the individuals in the group are all homogeneous with respect to the phenomena studied—thus, *ergodic*. In an ergodic system, all entities within the system are essentially interchangeable, such that knowledge of the entities' average scores

can be used to predict the scores of any of these entities. But given that—for psychology's study phenomena—ergodicity cannot be assumed, the common practice of aggregating data over individuals is equivalent to trying to find the mean of apples, pears and bananas. The performance of each individual of a group rarely, if ever, matches the groups' average performance—indeed, psychological variables are often optimised for representing normal distribution patterns in a group.

The *ergodic fallacy*—the practice of erroneously assuming that sample-level findings could inform about individual-level phenomena (Fisher et al., 2018; Molenaar and Campbell, 2009; Richters, 2021; Rose, 2016; Speelman and McGann, 2020)—can lead to erroneous interpretations of statistical test results. For instance, group differences in performance scores are commonly taken to indicate that "people" in one condition performed better than those in another—as if the difference between the two group means reflects a difference present in all, or at least most, of the individuals in the groups studied.

These problems are obscured by the ambiguous wording often used in conclusions. Speelman et al. (2024) analysed a year of articles ($N = 326$) from three highly cited Q1 journals in the fields of cognitive, educational and clinical psychology. Over 88% of the papers reported generic conclusions about "people" or "participants" when interpreting findings derived from group-level analysis (e.g., null-hypothesis significance tests). Prevalence of this error was highest in papers from cognitive psychology (93.3%), which typically assess claims about 'cognitive mechanisms' theorised as universal, compared to educational psychology (89.3%) and clinical psychology (77.9%), which are more concerned with individually relevant interventions. Still, prevalence of the ergodic fallacy was high in all fields.

## How the ergodic fallacy may influence psychology's reproducibility problems: Pervasiveness analysis as a suitable alternative to aggregationist statistics

The ergodic fallacy provides a straightforward explanation for reproducibility problems in psychology (Speelman and McGann, 2020). Without assessing whether an effect is pervasive, or even widely prevalent, in a given sample, it is difficult to know what to expect from replication. If a set of scores represents, for example, the idiosyncratic combination of individuals' idiosyncratic behaviours, then any attempt to reproduce an effect with another sample of individuals will involve a different set of scores that, however, likewise represent idiosyncratic combinations of idiosyncratic behaviours (Tang and Braver, 2020). Given this, it is unsurprising that many effects are difficult to replicate in psychological research (Iso-Ahola, 2024; Mayrhofer et al., 2024).

As a simple alternative to aggregationist statistical analysis methods, Speelman and McGann (2020) described *pervasiveness analysis*. This technique involves counting the number of individuals who exhibited a particular behaviour. Reaching a benchmark of 80% in a sample is considered sufficient evidence to support generic statements, such as "most individuals showed this behaviour under these circumstances". Moore et al. (2023) demonstrated the utility of this technique, by re-analysing the data of successful replications of nine famous psychology experiments, performed with null-hypothesis significance tests (Zwaan et al., 2018). Seven of these experiments met the pervasiveness criterion; that is, in each experiment, the target effect applied to over 80% of the participants. In the two other experiments, the classic effect applied to only 70% and 64% of the participants, respectively, although these experiments had passed the replicability criteria based on common significance tests.

Speelman and McGann's (2020) method for conducting a pervasiveness analysis is appropriate only for within-subjects designs. But pervasiveness analyses can also be applied to between-subject designs, correlational designs and forms of risk assessment. For these types of analyses, each set of findings is described in terms of "the number of persons who matched or failed to match expectation" (Grice et al., 2020, p.451) where the expectation is based on a theoretical prediction under test, such as more people given a drug will be classified as "cured" compared to people given a placebo. McManus et al. (2023, p. 2) extended this approach "to estimate the prevalence of person-level effects in the population" by comparing observed prevalence rates with null hypotheses of no effect. Interestingly, McManus and colleagues' re-analysis of existing data sets using this technique showed that previously reported statistically significant findings were often not associated with high pervasiveness values (also called prevalence values or Percent Correct Classification PCC indices). When surveying psychology researchers' knowledge of these problems, they also found that most researchers were largely ignorant of the potential dissociation between statistically significant effects and the pervasiveness of those effects in their samples.

Hence, pervasiveness analyses provide useful further insight into what is meant by an "effect" in a study and how many individuals of the sample actually met the desired criteria. They also showed how even successful replication studies can camouflage interesting and potentially important variation in (apparently) robust statistical outcomes. Importantly, though, pervasiveness analysis is unlikely to return a result of 100%—because of the non-ergodicity of human behaviour.

Pervasiveness analysis is an example of the epistemically justified analytical strategy that is necessary for generalising findings across unique individuals. Our next contribution elaborates on its methodological foundations and discusses suitable methodical approaches.

## Strategies for generalising findings across unique individuals in psychology: Misconceived nomotheticism and epistemically valid nomothetic approaches

As a science exploring individuals, psychology seems to contradict the old scientific dictum *scientia non est individuorum*[13]—the idea that scientific disciplines cannot be devoted to studying single cases given that science seeks regularities and lawfulness through abstraction and generalisation from particulars and unique events. Jana Uher explored the

13 Latin, meaning "science is not about individual cases".

epistemological and methodological fundamentals that can be derived from this dictum in her line of research on individuals within and across not just different human cultures but also different species (e.g., Uher, 2011, 2013, 2015a,c,d, 2018b,c, 2022b).

## Three strategies for generalising findings: Idiographic approaches, sample-based and case-by-case based nomothetic approaches

Windelband (1904/1998) categorised the sciences by their strategies of knowledge generation. Sciences of laws (e.g., physics, chemistry) study invariant relations of non-living matter (e.g., physical laws, chemical principles) using *nomothetic approaches* (from Greek *nomos*, the law). Sciences of events (e.g., history, sociology, political science), by contrast, study the ever-changing processes of human societies as they unfold through irreversible time using *idiographic approaches* (from Greek *idios*, the peculiar). Windelband's distinction reflects different strategies of knowledge generation that are aligned to the peculiarities of different objects of research. All sciences, however, apply both strategies—just to varied degrees because all research starts with a first case (Lamiell, 1998; Salvatore and Valsiner, 2010). Many sciences apply both strategies to equal extent. Evolutionary science, for example, studies unique events in the evolution of life (e.g., the dinosaurs' extinction) to derive general principles applicable to all species (e.g., adaptation, natural selection). Psychology, as well, studies unique individuals and aims to derive general principles that are applicable to many individuals. Thus, idiographic and nomothetic approaches are not mutually exclusive opposites, as often believed. Both are epistemically necessary and justified.

The physical sciences apply *sample-based nomothetic approaches* because (some of) their inanimate ergodic study systems feature synchronic and diachronic variations that are statistically isomorphic. Averages of many cases can therefore inform about every single case (e.g., electrons). To identify ('lawful'—nomothetic) regularities and universal principles in psychology, quantitative psychologists (e.g., Francis Galton) adopted this approach analogously (Lamiell, 2003). The majority uses sample-level analyses and generalises their findings to the single individuals thus-summarised (Figure 5). That is, individuals are studied only as abstract examples of prototypical—yet inexistent—individuals (Allport, 1937; Danziger, 1985b, 1990; Robinson, 2011). Sample-based nomothetic approaches have turned psychology into a science that is largely studying groups and populations rather than individuals—thus, into *psycho-demography* (Lamiell, 2018; Smedslund, 2021).

This also seriously limits psychologists' possibilities for causal analyses. Indeed, to group individuals, researchers must specify criteria (encoded as 'independent variables', e.g., gender, ethnicity) as possible causes of the phenomena analysed for between-group differences (e.g., intellectual abilities). These grouping criteria must be specified *a-priori*—thus, often before their relevance for a given research question is ascertained. For example, reviews of psychological meta-analyses showed that 78% of the effect sizes of reported gender differences were trivial or small (Cohen's $d < 0.2$; Hyde, 2005; Zell et al., 2015). Still, in the narrated interpretation, gender differences are often exaggerated, sometimes 'supported' by statistical significance levels, although these are known to depend on sample size. Analysing differences between researcher-defined groups often fails to generate findings that are informative about individuals' functioning and development and possible causally relevant differences between them (Danziger, 1990; Lamiell, 2003; Richters, 2021; Smedslund, 2016; Uher, 2015c, 2022b; van Geert, 2011). This is because sample-level nomothetic approaches disconnect theory development from descriptions of real individuals and cannot reveal what is, indeed, common to all individuals in a group.

To appropriately consider the peculiarities of psychology's study phenomena (e.g., non-ergodicity, higher-order complexity), alternative nomothetic approaches are required—and possible. In *case-by-case based nomothetic approaches,* which can be traced back to Wilhelm Wundt already (Lamiell, 2003), individuals are grouped by the commonalities and differences that they are shown to exhibit in the study phenomena (Figure 5). Considering many-to-one (degeneracy, equifinality) and one-to-many (pluripotency, multifinality) structure–function relations, the individuals within each of the thus-created groups are then explored for further commonalities and differences. For example, rather than analysing gender or ethnicity differences as a default, groups of individuals may be formed who are scoring low, medium vs. high in 'intelligence tests' to analyse what individuals within each group may have in common and what distinguishes them from those in the other groups, such as to identify possible factors promoting or hindering test performances. This nomothetic approach, because it is case-by-case based, allows researchers to identify generalities that are, indeed, common to all cases in a given group—a prerequisite for developing generalised knowledge and theories about intra-individual processes and functioning (Lamiell, 2003; Salvatore and Valsiner, 2010; Robinson, 2011; Uher, 2022b).

## Individual-/person-oriented rather than variable-oriented analyses

Empirical implementations of the two different nomothetic strategies are based on Stern's (1911) methodological framework

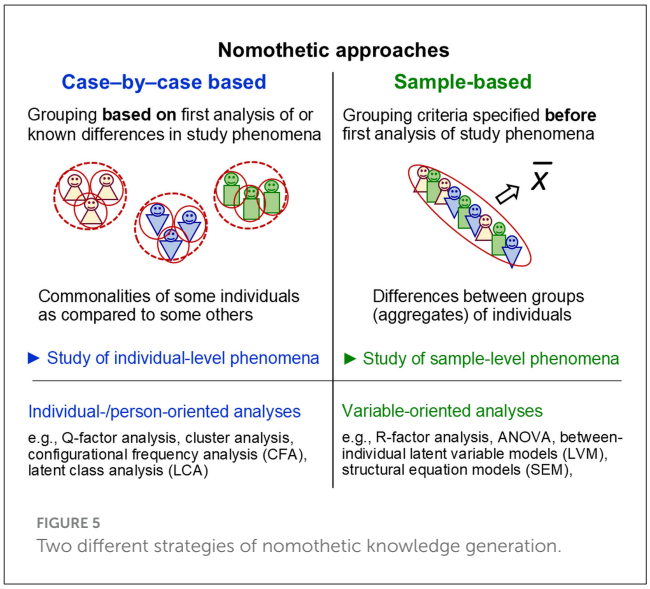



FIGURE 5
Two different strategies of nomothetic knowledge generation.

for exploring individuals and individual differences (Lamiell, 2003; Uher, 2011). It provides the necessary foundations for different, already well-established analytical methods to generalise findings across unique individuals.

Sample-based nomothetic approaches are empirically implemented through *variable-oriented analyses*, which explore the data matrix of $X_i$ individuals by $Y_j$ variables from the viewpoint of the $j$ variables to study their value distributions across all $i$ individuals. These methods analyse sample-level patterns in populations but not single individuals, such as using correlation or R factor analysis, ANOVA, between-individual latent variable models (LVMs) or structural equation models (SEM). Case-by-case based nomothetic approaches, by contrast, are empirically implemented through *individual-/person-oriented analyses*, which explore the data matrix from an orthogonal view and study the $i$ individuals for their value distributions across all $j$ variables. That is, these methods analyse *individual configurations* of values across different variables, which can be illustrated as a *profile* (e.g., 'intelligence' profile). This profile reflects a property of the individual, but not of the population. Individual-/person-oriented analyses can also be used to identify groups of individuals sharing similar configurations—thus, (profile) *types*—such as using Q factor analysis, configurational frequency analysis (CFA), latent class analysis (LCA) or cluster analysis (Bergman and Andersson, 2010; Bergman and Lundh, 2015; Bergman and Trost, 2006; Bergman et al., 2017; Lundh, 2023, 2024; Uher, 2011; von Eye and Bogat, 2006).

Individual-/person-oriented analyses allow researchers to scrutinise the implications of data aggregation as well as the limitations and possibilities of making inferences from groups (on different levels of aggregation) to single individuals, and vice versa (von Eye and Bergman, 2003). These methodological approaches underlie Grice's (2011) Observation-Oriented Modelling (OOM), Barrett's actuary approaches (Grice et al., 2017b) and Speelman and McGann's (2020) pervasiveness analysis. Weber's (2021) concepts of property scope and property order, in turn, are essential to conceptualise the non-ergodicity of psychology's study phenomena on ontological levels. These approaches and concepts are indispensable for exploring what is, in fact, common to all individuals of a group as an important prerequisite for tackling psychology's crisis in generalisability, replicability and validity.

## Conclusions and future directions: Psychology can no longer ignore its Questionable Research Fundamentals (QRFs)

In this article, we demonstrated that the currently discussed Questionable Research Practices (QRPs) are just surface-level symptoms that obscure the root causes of psychology's crises—its Questionable Research Fundamentals (QRFs) of many of its established (and therefore no longer questioned) theories, concepts, approaches, methods and practices (Figure 1). Our compilation of critical perspectives on psychology's crises and current issues pinpoints four major areas of future development to advance psychology's research fundamentals.

**(1) The systematic elaboration of psychology's general philosophy of science, especially of ontologies, epistemologies and methodologies**

We discussed different philosophy-of-science perspectives underlying the approaches that we critically analysed as well as those that we presented, highlighting their specific presuppositions as well as crucial differences between them. Our aim was to show (a selection of) the diversity of philosophies and theories of science that are being used in quantitative psychology. But our analyses also revealed Questionable Research Fundamentals (QRFs) in the form of contradictions and incompatibilities inherent in some widely-used approaches (e.g., in psychometrics), which preclude epistemically justified inferences on the phenomena studied. These serious issues often go unnoticed, however, because many psychologists follow established theories, methods and practices without scrutinising their philosophy-of-science fundamentals. To develop epistemically justified approaches, it is crucial to make the philosophical presuppositions on which a given line of research is built explicit, and thus accessible to analysis and elaboration. This is a prerequisite to establish coherent paradigms in which the specific ontology, epistemology and methodology used in a given line of research—no matter which specific ones may be preferred—are systematically aligned to one another.

**(2) The advancement of the philosophy-of-science fundamentals of specific theories, approaches and methods that are appropriate for enabling quantitative research considering the peculiarities of psychology's study phenomena**

We demonstrated Questionable Research Fundamentals (QRFs) also underlying common theories and approaches of psychological 'measurement' and pinpointed the challenges that must be mastered for establishing genuine analogues of measurement in psychology. To achieve this, quantitative psychologists must conceptualise how the peculiarities of its study phenomena (e.g., higher-order complexity, non-ergodicity) can be systematically connected to numerical (formal) models and known quantity standards. This also involves scrutinising the purported necessity and meaningfulness of quantitative investigations as well as the actual possibilities for implementing quantitative approaches and inevitable limitations.

**(3) The conceptual implementation of the epistemically necessary distinction between the phenomena under study and the means of their investigation**

Psychologists must heed the epistemic necessity to logically distinguish between the study phenomena (e.g., participants' beliefs, thoughts) and the means used for their exploration (e.g., methods, models) in a study in order to avoid conflating and thus confusing ontological with epistemological concepts (psychologists' cardinal error). This requires some basic knowledge about language and an increased awareness of its intricacies (e.g., inbuilt semantics). Such linguistic knowledge is necessary to explore and understand the challenges that these entail for psychological investigations, especially when using language-based methods (e.g., rating 'scales', item variables).

**(4) The establishment of epistemically justified strategies for generalising findings across unique individuals**

We demonstrated that psychology's default use of sample-based nomothetic approaches to study individual-level phenomena, implemented through statistical variable-oriented analyses, builds

on mathematical-statistical errors. It also ignores essential ontic peculiarities of its study phenomena, such as within-individual and between-individual variability, irreversible individual development and higher-order complexity (e.g., one-to-many and many-to-one relations, contextuality). These problems entail erroneous inferences from group-level findings to individual-level phenomena (e.g., ergodic fallacy), and vice versa, and also hinder causal analyses. To generalise across unique individuals, psychologists should capitalise on case-by-case based nomothetic approaches, implemented through individual-/person-oriented analyses for which the methodological fundamentals as well as suitable methods are already well established. These approaches are necessary to explore what some individuals do, in fact, have in common and what distinguishes them from others, which is prerequisite for unravelling (possibly) underlying structures and processes.

For each area of development, we presented various lines of research that, although established for years if not decades already, have still hardly been considered in mainstream psychology. With the increasing awareness of fundamental problems in psychological research and practice (e.g., psychology's crises), it is vital that more psychologists step out of their current comfort zone and start to actively and systematically advance the research fundamentals of psychological science. These novel directions can and should be built on the many fruitful developments that have already been made in psychology's history and diverse scientific communities. But these have been sidelined by the efficient mass production of purportedly 'quantitative' data through rating 'scales'. Their ease of use and efficiency enabled a blind *empiricism*—a focus on experience, largely disconnected from an elaborated body of theoretical knowledge—that fuelled the development of ever more sophisticated (and therefore impressive) statistical analyses—whereas psychology's actual study phenomena got out of focus.

## Just minimising Questionable Research Practices (QRPs) and using language-based algorithms will not remedy but only intensify psychology's crises

Mainstream psychologists launched large-scale initiatives (e.g., open science and replicability projects) to remedy questionable applications of established practices—thus, scientific misconduct. These approaches, however, encourage ever more empirical research—thus, mere empiricism—without elaborating the necessary theoretical and philosophical fundamentals. The novel technological possibilities provided by language-based algorithms (e.g., NLP algorithms, LLMs) allow for generating data sets even more rapidly than this has already been possible with the anonymous online surveys used in the last decades (Anderson et al., 2019)—and which are increasingly completed by online bots (Storozuk et al., 2020). The fascinating AI technologies have already generated an increasing volume of psychological research from artificially generated data to new ways of summarising findings. But this, in itself, will not address the serious issues underlying psychology's philosophies, theories and its language-based constructs and methods. Yet these novel technologies can be meaningfully applied to investigate how the inbuilt semantics of natural human languages mediate and shape individuals' thinking—including the theoretical thinking of scientists—and how individuals are relating their language to the real-world phenomena described.

## Psychology must tackle the Questionable Research Fundamentals (QRFs) of its established theories and practices and advance its philosophies of science

Tackling psychology's crises in replicability, generalisability, validity and confidence and the issues that cause and maintain them requires a rethinking of its established theories, methods and practices. Rather than trying to reinvent the wheel, mainstream psychology can and should capitalise on the advances already made over the last decades from different perspectives and fields of expertise. Therefore, we need more open and controversial yet constructive and collegial debates about our most basic presuppositions as well as honest and critical analyses of the possibilities and meaningfulness of quantification in psychology—*prioritising scientific integrity over expediency*. With our compilation of diverse perspectives on quantitative psychology's problems, we aim to set an example, to give new impetus to the current debates and to highlight important directions of future development that, as we believe, are necessary to rethink and advance psychology as a science.

## Author contributions

JU: Conceptualization, Project administration, Writing – original draft, Writing – review & editing, Visualization. JA: Writing – original draft, Visualization. PB: Writing – original draft. MH: Writing – original draft. J-HH: Writing – original draft. JM: Writing – original draft. LM: Writing – original draft. MM: Writing – original draft. RM: Writing – original draft. CS: Writing – original draft. AT: Writing – original draft. RW: Writing – original draft, Visualization.

## Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

## Acknowledgments

We thank Robert Barrett (PhD) for language editing Paul's final draft. We are also grateful to Bob's two daughters as well as to Randy E. Bennett and Jiangang Hao for reviewing Bob's final draft.

## In memoriam

We dedicate this article to our dear colleagues and co-authors Paul Barrett and Robert J. Mislevy who died before they could see this article in print. Paul's tireless efforts to promote and argue

for scientific integrity and a "think-for-yourself" attitude as well as Bob's integrative efforts for advancing the philosophy, theory and practice of educational assessment will be sorely missed.

## Conflict of interest

PB was employed by Advanced Projects R&D Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

## Publisher's note



## References

Abran, A., Desharnais, J.-M., and Cuadrado-Gallego, J. J. (2012). Measurement and quantification are not the same: ISO 15939 and ISO 9126. *J. Softw. Evol. Process* 24, 585–601. doi: 10.1002/smr.496

Adam, M., and Hanna, P. (2012). Your past is not their present: time, the other, and ethnocentrism in cross-cultural personality psychology. *Theory Psychol.* 22, 436–451. doi: 10.1177/0959354311412107

AERA, APA, and NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Aeschliman, M. D. (1998). *The Restitution of Man: C. S. Lewis and the Case Against Scientism*. Grand Rapids, MI: William B. Eerdmans Publishing Company.

Al-Ababneh, M. (2020). Linking ontology, epistemology and research methodology. *Sci. Philos.* 8, 75–91. doi: 10.23756/sp.v8i1.500

Ali, M. (2023). "Research philosophies in social science and information systems research," in *Information Systems Research* (Cham: Palgrave Macmillan, Cham), 256.

Allport, G. W. (1937). *Personality: A Psychological Interpretation*. New York, NY: Holt, Rinehart & Winston.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., and Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Pers. Soc. Psychol. Bull.* 45, 842–850. doi: 10.1177/0146167218798821

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *J. Clin. Psychiatry* 82:20f13804. doi: 10.4088/JCP.20f13804

Arnulf, J. K., and Larsen, K. R. (2020). Culture blind leadership research: how semantically determined survey data may fail to detect cultural differences. *Front. Psychol.* 11:176. doi: 10.3389/fpsyg.2020.00176

Arnulf, J. K., and Larsen, K. R. (2021). "Semantic and ontological structures of psychological attributes," in *Measuring and Modeling Persons and Situations*, eds. D. Wood, S. J. Read, P. D. Harms, and A. Slaughter (London: Academic Press), 69–102. doi: 10.1016/B978-0-12-819200-9.00013-2

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Bong, C. H. (2014). Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS One* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Egeland, T. (2018). The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength. *Behav. Res. Methods* 50, 2345–2365. doi: 10.3758/s13428-017-0999-y

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Nimon, K. F. (2021). Editorial: semantic algorithms in the assessment of attitudes and personality. *Front. Psychol.* 12:720559. doi: 10.3389/fpsyg.2021.720559

Arnulf, J. K., Olsson, U. H., and Nimon, K. (2024). Measuring the menu, not the food: "psychometric" data may instead measure "lingometrics" (and miss its greatest potential). *Front. Psychol.* 15:1308098. doi: 10.3389/fpsyg.2024.1308098

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: methodological and philosophical foundations. *MIS Q.* 35, 261–292. doi: 10.2307/23044044

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.* 84:191. doi: 10.1037/0033-295X.84.2.191

Barrett, P. T. (2003). Beyond psychometrics: measurement, non-quantitative structure, and applied numerics. *J. Manag. Psychol.* 18, 421–439. doi: 10.1108/02683940310484026

Barrett, P. T. (2005). What if there were no psychometrics?: constructs, complexity, and measurement. *J. Pers. Assess.* 85, 134–140. doi: 10.1207/s15327752jpa8502_05

Barrett, P. T. (2008). The consequence of sustaining a pathology: scientific stagnation—a commentary on the target article "Is psychometrics a pathological science?" by Joel Michell. *Meas. Interdiscip. Res. Perspect.* 6, 78–83. doi: 10.1080/15366360802035521

Barrett, P. T. (2011). Invoking arbitrary units is not a solution to the problem of quantification in the social sciences. *Meas. Interdiscip. Res. Perspect.* 9, 28–31. doi: 10.1080/15366367.2011.558783

Barrett, P. T. (2018). The EFPA test-review model: when good intentions meet a methodological thought disorder. *Behav. Sci.* 8, 1 5, 1–22. doi: 10.3390/bs8010005

Barrett, P. T. (2024). *The Cognadev AI-Series #1-4*. Available online at: https://www.cognadev.com/blog_cat11.html (Accessed October 30, 2024).

Berglund, B. (2012). "Measurement in psychology," in *Measurement with Persons: Theory, Methods, and Implementation Areas,* eds. B. Berglund, G. B. Rossi, J. T. Townsend, and L. Pendrill (New York: Taylor Francis), 27–50.

Bergman, L. R., and Andersson, H. (2010). The person and the variable in developmental psychology. *J. Psychol.* 218, 155–165. doi: 10.1027/0044-3409/a000025

Bergman, L. R., and Lundh, L.-G. (2015). The person-oriented approach: roots and roads to the future. *J. Person-Oriented Res.* 1, 1–109. doi: 10.17505/jpor.2015.01

Bergman, L. R., and Trost, K. (2006). The person-oriented versus the variable-oriented approach: are they complementary, opposites, or exploring different worlds? *Merrill-Palmer Q.* 52, 601–632. doi: 10.1353/mpq.2006.0023

Bergman, L. R., Vargha, A., and Kövi, Z. (2017). Revitalizing the typological approach: some methods for finding types. *J.Person Orient. Res.* 3, 49–62. doi: 10.17505/jpor.2017.04

Bernstein, J. H. (2015). Transdisciplinarity: a review of its origins, development, and current issues. *J. Res. Pract.* 11:412. Available online at: https://jrp.icaap.org/index.php/jrp/article/view/510.html (Accessed September 23, 2019).

Bevir, M., and Blakely, J. (2018). *Interpretive Social Science: An Anti-Naturalist Approach*. Oxford: Oxford University Press.

Bhaskar, R., and Danermark, B. (2006). Metatheory, interdisciplinarity and disability research: a critical realist perspective. *Scand. J. Disabil. Res.* 8, 278–297. doi: 10.1080/15017410600914329

Bickhard, M. H. (2001). The tragedy of operationalism. *Theory Psychol.* 11, 35–44. doi: 10.1177/0959354301111002

Bird, A. (2022). "Thomas Kuhn", in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/spr2022/entries/thomas-kuhn (Accessed May 23, 2025).

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. U.S.A.* 17, 656–660. doi: 10.1073/pnas.17.2.656

Bollen, K. A., and Diamantopoulos, A. (2017). In defense of causal-formative indicators: a minority report. *Psychol. Methods* 22, 581–596. doi: 10.1037/met0000056

Boroditsky, L. (2018). *7,000* universes: how the languages we speak shape the ways we think. Toronto, ON: Doubleday Canada.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika* 71, 425–440. doi: 10.1007/s11336-006-1447-6

Borsboom, D. (2008). Latent variable theory. *Meas. Interdiscip. Res. Perspect.* 6, 25–53. doi: 10.1080/15366360802035497

Borsboom, D., and Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory Psychol.* 14, 105–120. doi: 10.1177/0959354304040200

Borsboom, D., and Scholten, A. Z. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory Psychol.* 18, 111–117. doi: 10.1177/0959354307086925

Bridgman, P. W. (1927). *The Logic of Modern Physics*. New York: Macmillan.

Bridgman, P. W. (1938). Operational analysis. *Philos. Sci.* 5, 114–131. doi: 10.1086/286496

Bruner, J. (1990). *Acts of Meaning*. Cambridge: Harvard University Press.

Bunge, M. (1977). *Treatise on Basic Philosophy: Ontology I: The Furniture of the World*, Vol. 3. Dordrecht, Boston: D. Reidel Publishing.

Bunge, M. (1981). "A materialist theory of mind," in *Scientific Materialism. Episteme*, Vol. 9 (Dordrecht: Springer), 67–89.

Bunge, M. (1993). Realism and antirealism in social science. *Theory Decis.* 35, 207–235. doi: 10.1007/BF01075199

Bunge, M. A. (1974). *Treatise on Basic Philosophy Volume 1: Semantics I – Sense and Reference*. Dordrecht, The Netherlands: D. Reidel Publishing Company.

Bunge, M. A. (1975). What is a quality of life indicator? *Soc. Indic. Res.* 2, 65–79. doi: 10.1007/BF00300471

Bunge, M. A. (1979). *Treatise on Basic Philosophy Volume 4: Ontology II - A World of Systems*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bunge, M. A. (2003). *Philosophical Dictionary (Enlarged ed.)*. Amherst, New York: Prometeus Books.

Bunge, M. A. (2010). Reading measuring instruments. *Spont. Generations* 4, 85–93. doi: 10.4245/sponge.v4i1.11725

Byrne, D. (2002). *Interpreting Quantitative Data*. London: Sage Publications.

Campbell, D. T. (1974). "Qualitative knowing in action research. Kurt Lewin award address," in *Society for the Psychological Study of Social Issues, Presented at the Meeting of the American Psychological Association* (New Orleans, LA: APA).

Chakravartty, A. (2017). "Scientific realism", in *The Stanford Encyclopedia of Philosophy (Summer 2017 Edition),* ed. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/sum2017/entries/scientific-realism (Accessed May 13, 2025).

Chang, H. (1995). Circularity and reliability in measurement. *Perspect. Sci.* 3, 153–172. doi: 10.1162/posc_a_00479

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15:39. doi: 10.1145/3641289

Cicourel, A. V. (1964). *Method and Measurement in Sociology*. New York, NY: Free Press of Glencoe.

Collingwood, R. G. (1940). *An Essay on Metaphysics*. Oxford: Oxford University Press.

Cordero, A. (2012). Mario Bunge's scientific realism. *Sci. Educ.* 21, 1419–1435. doi: 10.1007/s11191-012-9456-6

Cornejo, C., and Valsiner, J. (2021). "Mathematical thinking, social practices, and the locus of science in psychology," in *A Pragmatic Perspective of Measurement* (S. vii–xi), eds. D. Torres Irribarra (Cham: Springer International Publishing).

Danziger, K. (1985a). The origins of the psychological experiment as a social institution. *Am. Psychol.* 40, 133–140. doi: 10.1037//0003-066X.40.2.133

Danziger, K. (1985b). The methodological imperative in psychology. *Philos. Soc. Sci.* 15, 1–13. doi: 10.1177/004839318501500101

Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*. New York, NY: Cambridge University Press.

Danziger, K., and Dzinas, K. (1997). How psychology got its variables. *Can. Psychol.* 38, 43–48. doi: 10.1037/0708-5591.38.1.43

Daston, L., and Galison, P. (2007). *Objectivity*. New York, NY: Zone Books.

de Lubac, H. (1995). *The Drama of Atheist Humanism*. San Francisco, CA: Ignatius Press.

Deutscher, G. (2006). *The Unfolding of Language: The Evolution of Mankind's Greatest Invention*. London, UK: Arrow.

Deutscher, G. (2010). *Through the language glass: Why the World Looks Different in Other Languages*. New York, NY: Metropolitan Books/Henry Holt and Company.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* [Preprint]. *arXiv:*1810.04805. doi: 10.48550/arXiv.1810.04805

Diez Roux, A. V. (2002). A glossary for multilevel analysis. *J. Epidemiol. Commun. Health* 56, 588–594. doi: 10.1136/jech.56.8.588

Durkheim, E. (1893/1984). *The Division of Labor in Society (W. D. Halls, Trans.).* New York, NY: Free Press.

Earp, B. D., and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Front. Psychol.* 6:621. doi: 10.3389/fpsyg.2015.00621

Edwards, J. R. (2011). The fallacy of formative measurement. *Organ. Res. Methods* 14, 370–388. doi: 10.1177/1094428110378369

Elson, M., Hussey, I., Alsalti, T., and Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Commun. Psychol.* 1, 1–4. doi: 10.1038/s44271-023-00026-9

Eronen, M. I. (2024). Causal complexity and psychological measurement. *Philos. Psychol.* 1–16. doi: 10.1080/09515089.2023.2300693

Fahrenberg, J. (2013). *Zur Kategorienlehre der Psychologie: Komplementaritätsprinzip; Perspektiven und Perspektiven-Wechsel [On Category Systems in Psychology*. Complementarity Principle. Perspectivism and Perspective-taking]. Lengerich, Germany: Pabst Science Publishers.

Fahrenberg, J. (2015). *Theoretische Psychologie – Eine Systematik der Kontroversen. [Theoretical psychology – A systematization of controversies]*. Lengerich, Germany: Pabst Science Publishers.

Faucheux, C. (1976). Cross-cultural research in experimental social psychology. *Eur. J. Soc. Psychol.* 6, 269–322. doi: 10.1002/ejsp.2420060302

Faust, D. (2012). *Ziskin's Coping with Psychiatric and Psychological Testimony*. New York: Oxford University Press.

Fechner, G. T. (1858). Das Psychische Maß. *Z. Philos. Kritik* 32, 1–24.

Fechner, G. T. (1860a). *Elemente der Psychophysik I* (Vol. 1). Leipzig: Breitkopf und Härtel.

Fechner, G. T. (1860b). *Elemente der Psychophysik II* (Vol. 2). Leipzig: Breitkopf und Härtel.

Feest, U. (2005). Operationism in psychology: what the debate is about, what the debate should be about. *J. Hist. Behav. Sci.* 41, 131–149. doi: 10.1002/jhbs.20079

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., et al. (1940). Quantitative estimates of sensory events: final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Adv. Sci.* 1, 331–349.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement* 34, 39–48. doi: 10.1016/S0263-2241(03)00018-6

Fisher, A. J., Medaglia, J. D., and Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proc. Natl. Acad. Sci. USA.* 115, E6106–E6115. doi: 10.1073/pnas.1711978115

Fleck, L. (1935/1979). *Genesis and development of a scientific fact [Translation of: Entstehung und Entwicklung einer wissenschaftlichen Tatsache, original 1935]*. Chicago: University of Chicago Press.

Flynn, J. R. (2012). *Are We Getting Smarter? Rising IQ in the Twenty-First Century*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139235679

Freeman, M. (2024). *Toward the Psychological Humanities. A Modest Manifesto for the Future of Psychology*. London: Routledge.

Frigg, R., and Nguyen, J. (2021). "Scientific representation," in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta (Cambridge: Cambridge University Press). Available online at: https://plato.stanford.edu/archives/win2021/entries/scientific-representation (Accessed March 31, 2025).

Gee, J. P. (2008). "Learning theory, video games, and popular culture," in *The International Handbook of Children, Media, and Culture,* eds. K. Drotner and S. Livingstone (London: Sage Publications), 196–211.

Gee, J. P. (2021). "Thinking, learning, and reading: the situated social mind," in *Situated Cognition: Social, Semiotic, and Psychological Perspectives,* eds. D. Kirshner and Whitson (New York: Routledge), 235–259.

Gergen, K. J. (1973). Social psychology as history. *J. Pers. Soc. Psychol.* 26, 309–320. doi: 10.1037/h0034436

Gergen, K. J. (2001). Psychological science in a postmodern context. *Am. Psychol.* 56, 803–813. doi: 10.1037/0003-066X.56.10.803

Gibbs, P., and Beavis, A. (2020). *Contemporary Thinking on Transdisciplinary Knowledge: What Those Who Know, Know*. Cham, Switzerland: Springer.

Gnatt, E. E. (2018). "Scientism and saturation: evolutionary psychology, human experience, and the phenomenology of Jean-Luc Marion," in *On Hijacking Science. Exploring the Nature of Consequences of Overreach in Pychology,* eds. E. E. Gnatt and R. N. Williams (New York, NY: Routledge), 52–67.

Gong, T., Shuai, L., and Mislevy, R. J. (2023). Sociocognitive processes and item response models: a didactic example. *J. Educ. Meas.* 61, 150–173. doi: 10.1111/jedm.12376

Gould, S. J. (1996). *The Mismeasure of Man (extended and revised ed.).* New York: W.W. Norton.

Gray, R. M. (1988). "Ergodic theorems," in *Probability, Random Processes, and Ergodic Properties* (New York, NY: Springer), 216–243.

Green, C. D. (2001). Operationism again: what did Bridgman say? What did Bridgman need? *Theory Psychol.* 11, 45–51. doi: 10.1177/0959354301111003

Grice, J. (2011). *Observation Oriented Modeling: Analysis of Cause in the Behavioral Sciences.* New York: Academic Press.

Grice, J., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., et al. (2017b). Four bad habits of modern psychologists. *Behav. Sci.* 7:53. doi: 10.3390/bs7030053

Grice, J., Barrett, P., Schlimgen, and Abramson, C. (2012). Toward a brighter future for psychology as an observa/on oriented science. *Behav. Sci.* 2, 1–22. doi: 10.3390/bs2010001

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., et al. (2020). Persons as effect sizes. *Adv. Methods Pract. Psychol. Sci.* 3, 443–455. doi: 10.1177/2515245920922982

Grice, J. W., Yepez, M., Wilson, N. L., and Shoda, Y. (2017a). Observation-oriented modeling: going beyond "is it all a matter of chance"? *Educ. Psychol. Meas.* 77, 855–867. doi: 10.1177/0013164416667985

Guyon, H. (2018). The fallacy of the theoretical meaning of formative constructs. *Front. Psychol.* 9:179. doi: 10.3389/fpsyg.2018.00179

Hacking, I. (1990). *The Taming of Chance*. New York: Cambridge University Press.

Hacking, I. (2002). *Historical Ontology.* Cambridge, MA: Harvard University Press.

Hammack, P. L., and Josselson, R. (2021). *Essentials of Narrative Analysis*. Washington, DC: American Psychological Association.

Hanfstingl, B., Oberleiter, S., Pietschnig, J., Tran, U. S., and Voracek, M. (2024). Detecting jingle and jangle fallacies by identifying consistencies and variabilities in study specifications – a call for research. *Front. Psychol.* 15:1404060. doi: 10.3389/fpsyg.2024.1404060

Hardin, A. M., and Marcoulides, G. A. (2011). A commentary on the use of formative measurement. *Educ. Psychol. Meas.* 71, 753–764. doi: 10.1177/0013164411414270

Harré, R., and Van Langenhove, L. (1999). *Positioning Theory: Moral Contexts of Intentional Action.* Oxford: Blackwell.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., and McElhaney, K. W. (2016). *Constructing Assessment Tasks that Blend Disciplinary Core Ideas, Crosscutting Concepts, and Science Practices for Classroom Formative Applications*. Menlo Park, CA: SRI International.

Hartmann, N. (1964). *Der Aufbau der realen welt. Grundriss der allgemeinen Kategorienlehre [the Structure of the Real World. Outline of the General Theory of Categories],* 3rd Edn. Berlin: Walter de Gruyter.

Harvard, S., and Winsberg, E. (2022). The epistemic risk in representation. *Kennedy Inst. Ethics J.* 31, 1–31. doi: 10.1353/ken.2022.0001

Heene, M. (2011). An old problem with a new solution, raising classical questions: a commentary on Humphry. *Measurement* 9, 51–54. doi: 10.1080/15366367.2011.558790

Heine, J. H., and Heene, M. (2025). Measurement and mind: unveiling the self-delusion of metrification in psychology. *Meas. Interdiscip. Res. Perspect.* 23, 213–241. doi: 10.1080/15366367.2024.2329958

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik [On The Actual Content of Quantum Theoretical Kinematics and Mechanics]. *Zeit. Phys.* 43, 172–198. doi: 10.1007/BF01397280

Hempel, C. G. (1952). "Fundamentals of concept formation in empirical science," in *International Encyclopedia of Unified Science,* Vol. 2, ed. H. C. Gustav (Chicago: University of Chicago Press). Available online at: https://archive.org/details/fundamentalsofco0000hemp (Accessed November 28, 2024).

Hibberd, F. J. (2019). What is scientific definition? *J. Mind Behav.* 40, 29–52. Available online at: https://www.jstor.org/stable/26740746

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe* 53, 1–64.

Holzkamp, K. (1983). *Grundlegung der Psychologie*. Frankfurt/M.: Campus.

Holzkamp, K. (2013). "Missing the point: variable psychology's blindness to the problem's inherent coherences," in *Psychology from the standpoint of the subject: Selected writings of Klaus Holzkamp*, eds. E. Schraube and U. Osterkamp (London: Palgrave Macmillan), 60–74.

Howell, K. E. (2013). *An Introduction to the Philosophy of Methodology*. London: SAGE Publications Ltd.

Hyde, J. S. (2005). The gender similarities hypothesis. *Am. Psychol.* 60, 581–592. doi: 10.1037/0003-066X.60.6.581

Ichheiser, G. (1943). Why psychologists tend to overlook certain "obvious" facts. *Philos. Sci.* 10, 204–207. doi: 10.1086/286811

Iso-Ahola, S. E. (2024). Science of psychological phenomena and their testing. *Am. Psychol.* doi: 10.1037/amp0001362. [Epub ahead of print].

James, W. (1895). The knowing of things together. *Psychol. Rev.* 2, 105–124. doi: 10.1037/h0073221

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Jovanović, G. (2022). "Epistemology of psychology," in *International Handbook of Psychology Learning and Teaching, Springer International Handbooks of Education,* eds. J. Zumbach, D. Bernstein, S. Narciss, and G. Marsico (Cham, Switzerland: Springer, Cham).

Kane, M. T. (1992). An argument-based approach to validity. *Psychol. Bull.* 112, 527–535. doi: 10.1037/0033-2909.112.3.527

Kaplan, A. (1964). *The Conduct of Inquiry: Methodology for Behavioral Science*. Scranton, PA: Chandler Publishing Co.

Kelley, T. L. (1927). *Interpretation of Educational Measurements*. Yonkers, NY: World.

Kelly, G. (1955). *The Psychology of Personal Constructs (Volume 1 and 2)*. London, UK: Routledge.

Kelly, G. (1963). *A Theory of Personality: The Psychology of Personal Constructs*. London: W.W. Norton.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior (Complex adaptive systems).* Cambridge, Mass: The MIT Press.

Kirschenbaum, H. (2007). *The Life and Work of Carl Rogers.* Alexandria, VA: PCCS Books.

Kirschner, S. R., and Martin, J. (2010). *The Sociocultural Turn in psychology: The Contextual Emergence of Mind and Self.* New York: Columbia University Press.

Klima, G, (2022). "The medieval problem of universals", in *The Stanford Encyclopedia of Philosophy (Spring 2022 Edition),* ed. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/spr2022/entries/universals-medieval (Accessed May 14, 2025).

Koch, S. (1992). Psychology's Bridgman vs Bridgman's Bridgman: an essay in reconstruction. *Theory Psychol.* 2, 261–290. doi: 10.1177/0959354392023002

Koffka, K. (1935). *Principles of Gestalt Psychology.* London: Routledge and Kegan Paul.

Krantz, D., Luce, R. D., Tversky, A., and Suppes, P. (1971). *Foundations of Measurement Vol. I: Additive and Polynomial Representations.* San Diego: Academic Press.

Kuhn, T. (1962/1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press (1970, 2nd edition, with postscript).

Lamiell, J. T. (1998). 'Nomothetic' and 'Idiographic': contrasting Windelband's understanding with contemporary usage. *Theory Psychol.* 8, 23–38. doi: 10.1177/0959354398081002

Lamiell, J. T. (2003). *Beyond Individual and Group Differences: Human Individuality, Scientific Psychology, and William Stern's Critical Personalism*. Thousand Oaks, CA: Sage Publications.

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: what is it? How was it contracted? Is there a cure? *New Ideas Psychol.* 31, 65–71. doi: 10.1016/j.newideapsych.2011.02.009

Lamiell, J. T. (2018). "On scientism in psychology: some observations of historical relevance," in *On Hijacking Science. Exploring the Nature of Consequences of Overreach in Psychology,* eds. E. E. Gnatt and R. N. Williams (New York, NY: Routledge), 27–41.

Lamiell, J. T. (2019a). *Psychology's Misuse of Statistics and Persistent Dismissal of its Critics*. Cham: Springer International. doi: 10.1007/978-3-030-12131-0

Lamiell, J. T. (2019b). Re-centering psychology: from variables and statistics to persons and their stories. *Theory Psychol.* 29, 282–284. doi: 10.1177/0959354318766714

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.*104, 211–240. doi: 10.1037/0033-295X.104.2.211

Larsen, K. R., Voronovich, Z. A., Cook, P. F., and Pedro, L. W. (2013). Addicted to constructs: Science in reverse? *Addiction* 108, 1532–1533. doi: 10.1111/add.12227

Legg, C., and Hookway, C. (2024). "Pragmatism", in *The Stanford Encyclopedia of Philosophy*, eds. E. N. Zalta and U. Nodelman. Available online at: https://plato.stanford.edu/archives/win2024/entries/pragmatism (Accessed June 03,2025).

Lewin, K. (1936). *Principles of Topological Psychology*. New York, NY: McGraw-Hill.

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 22, 1–55.

Linkov, V. (2024). Qualitative (pure) mathematics as an alternative to measurement. *Front. Psychol.* 15:1374308. doi: 10.3389/fpsyg.2024.1374308

Lord, F. M. (2012). *Applications of Item Response Theory to Practical Testing Problems*. New York, NY: Routledge.

Lovasz, N., and Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. *New Ideas Psychol.* 31, 22–31. doi: 10.1016/j.newideapsych.2011.02.005

Luce, R. D., Krantz, D., Suppes, P., and Tversky, A. (1990). *Foundations of Measurement, Vol. 3: Representation, Axiomatization, and Invariance*. San Diego, CA: Academic Press.

Luchetti, M. (2020). From successful measurement to the birth of a law: disentangling coordination in Ohm's scientific practice. *Stud. Hist. Philos. Sci. Part A* 84, 119–131. doi: 10.1016/j.shpsa.2020.09.005

Luchetti, M. (2024). Epistemic circularity and measurement validity in quantitative psychology: insights from Fechner's psychophysics. *Front. Psychol.* 15:1354392. doi: 10.3389/fpsyg.2024.1354392

Lundh, L. G. (2023). Person, population, mechanism. Three main branches of psychological science. *J. Person Orient. Res.* 9, 75–92. doi: 10.17505/jpor.2023.25814

Lundh, L. G. (2024). Person, population, mechanism. A rejoinder to critics and an elaboration of the three-branch model. *J. Person Orient. Res.* 10, 68–84. doi: 10.17505/jpor.2024.26295

Luria, A. R. (1973). *The Working Brain. An Introduction to Neuropsychology*. New York: Basic Books.

MacKenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Q.* 35, 293–334. doi: 10.2307/23044045

Madsen, O. J. (2014). *The Therapeutic Turn: How Psychology Altered Western Culture*. Hove, East Sussex: Routledge.

Mahner, M. (2021). Mario Bunge (1919–2020): conjoining philosophy of science and scientific philosophy. *J. Gen. Philos. Sci.* 52, 3–23. doi: 10.1007/s10838-021-09553-7

Maraun, M. D. (1998). Measurement as a normative practice: implications of Wittgenstein's philosophy for measurement in psychology. *Theory Psychol.* 8, 435–461. doi: 10.1177/0959354398084001

Maraun, M. D., and Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas Psychol.* 31, 32–42. doi: 10.1016/j.newideapsych.2011.02.006

Maraun, M. D., and Halpin, P. F. (2008). Manifest and latent variates. *Measurement* 6, 113–117. doi: 10.1080/15366360802035596

Margenau, H. (1950). *The Nature of Physical Reality. A Philosophy of Modern Physics*. New York: McGraw-Hill.

Mari, L., Carbone, P., Giordani, A., and Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Stud. Hist. Philos. Sci.* 65–66, 46–56. doi: 10.1016/j.shpsa.2017.08.001

Mari, L., Carbone, P., and Petri, D. (2015). "Fundamentals of hard and soft measurement," in *Modern Measurements: Fundamentals and Applications,* eds. A. Ferrero, D. Petri, P. Carbone, and M. *Catelani* (Hoboken, NJ: John Wiley and Sons), 203–262.

Mari, L., Maul, A., Irribarra, D. T., and Wilson, M. (2013). Quantification is neither necessary nor sufficient for measurement. *J. Phys. Conf. Ser.* 459:012007. doi: 10.1088/1742-6596/459/1/012007

Mari, L., Wilson, M., and Maul, A. (2021). "Measurement across the sciences," in *Developing a Shared Concept System For Measurement* (Cham: Springer). doi: 10.1007/978-3-030-65558-7

Markus, K. A. (2021). Philosophical methodology and axiomatic measurement theory: a comment on Uher (2021). *J. Theor. Philos. Psychol.* 41, 85–90. doi: 10.1037/teo0000178

Martin, J. (2013). Life positioning analysis: an analytic framework for the study of lives and life narratives. *J. Theor. Philos. Psychol.* 33, 1–17. doi: 10.1037/a0025916

Martin, J. (2017). Carl Rogers' and B. F. Skinner's approaches to personal and societal improvement: a study in the psychological humanities. *J. Theor. Philos. Psychol.* 37, 214–229. doi: 10.1037/teo0000072

Martin, J. (2022). "A non-reductive "person-based ontology" for psychological inquiry," in *Routledge International Handbook of Theoretical and Philosophical Psychology*, eds. B. D. Slife, S. C. Yanchar, and F. C. Richardson (New York, NY: Routledge), 391–411.

Martin, J. (2024). *Studies of Life Positioning: A New Sociocultural Approach to Psychobiography*. New York, NY: Routledge.

Martin, J., and Bickhard, M. H. (Eds.) (2013). *The Psychology of Personhood: Philosophical, Historical, Socio-Developmental, and Narrative Approaches*. New York: Columbia University Press.

Martin, J., and McLellan, A. (2013). *The Education of Selves: How Psychology Transformed Students.* New York: Oxford University Press.

Martin, J., and Sugarman, J. (1999). *The Psychology of Human Possibility and Constraint*. Albany, NY: State University of New York Press.

Martin, J., and Sugarman, J. (2009). Does interpretation in psychology differ from interpretation in natural science? *J. Theory Soc. Behav.* 39, 19–37. doi: 10.1111/j.1468-5914.2008.00394.x

Martin, J., Sugarman, J., and Hickinbottom, S. (2010). *Persons: Understanding Psychological Selfhood and Agency*. New York: Springer.

Martin, J., Sugarman, J., and Slaney, K. L. (Eds.) (2015). *Wiley Handbook of Theoretical and Philosophical Psychology: Methods, Approaches and New Directions for Social Science.* Wiley Blackwell.

Martin, J., Sugarman, J., and Thompson, J. (2003). *Psychology and the Question of Agency*. Albany: State University of New York Press.

Maslow, A. H. (1966). *The Psychology of Science.* A reconnaissance. New York: Gateway.

Mason, P. H. (2010). Degeneracy at multiple levels of complexity. *Biol. Theory* 5, 277–288. doi: 10.1162/BIOT_a_00041

Mason, P. H. (2015). Degeneracy: demystifying and destigmatizing a core concept in systems biology. *Complexity* 20, 12–21. doi: 10.1002/cplx.21534

Matthews, M. R. (Ed.) (2019). *Mario Bunge: A Centenary Festschrift*. Cham: Springer.

Mayrhofer, R., Büchner, I. C., and Hevesi, J. (2024). The quantitative paradigm and the nature of the human mind. The replication crisis as an epistemological crisis of quantitative psychology in view of the ontic nature of the psyche. *Front. Psychol.* 15:1390233. doi: 10.3389/fpsyg.2024.1390233

Mazur, L. B. (2015). Defining power in social psychology. *Orbis Idearum* 2, 101–114. doi: 10.26106/0D0Q-H908

Mazur, L. B. (2017). "Gaps in human knowledge: highlighting the whole beyond our conceptual reach," in *The Psychology of Imagination: History, Theory and New Research Horizons,* eds. B. Wagoner, I. Bresco, and S. H. Awad (Charlotte: Information Age Publishing), 239–252.

Mazur, L. B. (2021). The epistemological imperialism of science. Reinvigorating early critiques of scientism. *Front. Psychol.* 11:609823. doi: 10.3389/fpsyg.2020.609823

Mazur, L. B. (2022). "Experimentation within the social identity approach. History, highlights, and hurdles," in *Cambridge Handbook of Identity*, eds. M. Bamberg, C. Demuth, and M. Watzlawik (Cambridge: Cambridge University Press), 435–459.

Mazur, L. B. (2024a). "A dim recognition." Religion as a font of psychological innovation. *Integr. Psychol. Behav. Sci.* 58, 845–854. doi: 10.1007/s12124-024-09858-4

Mazur, L. B. (2024b). The desire for power within activist burnout. An illustration of the value of interpretive social science. *Sociol. Compass* 18:e13186. doi: 10.1111/soc4.13186

Mazur, L. B. (2024c). "The Victim" and the democratization of victimhood. *J. Theor. Philos. Psychol.* doi: 10.1037/teo0000290. [Epub ahead of print].

Mazur, L. B., Richter, L., Manz, P., and Bartels, H. (2022). The importance of cultural psychological perspectives in pain research: towards the palliation of Cartesian anxiety. *Theory Psychol.* 32, 183–201. doi: 10.1177/09593543211059124

Mazur, L. B., and Sticksel, I. (2021). An empirical study of psychology and logic. Abduction and belief as normalizing habits of positive expectation. *New Ideas Psychol.* 63:100874. doi: 10.1016/j.newideapsych.2021.100874

Mazur, L. B., and Watzlawik, M. (2016). Debates about the scientific status of psychology: looking at the bright side. *Integr. Psychol. Behav. Sci.* 50, 555–567. doi: 10.1007/s12124-016-9352-8

McGrane, J. (2015). Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Front. Psychol.* 6:431. doi: 10.3389/fpsyg.2015.00431

McManus R. M., Young, L., and Sweetman, J. (2023). Psychology is a property of persons, not averages or distributions: confronting the group-to-person generalizability problem in experimental psychology. *Adv. Methods Pract. Psychol. Sci.* 6:1186615. doi: 10.1177/25152459231186615

Mertens, D. M. (2023). *Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods,* 6th Edn. Thousand Oaks, CA: SAGE.

Messick, S. (1989). "Validity," in *Educational Measurement,* 3rd Edn., ed. R. L. Linn (New York: American Council on Education/Macmillan), 13–103.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *Br. J. Psychol.* 88, 355–383. doi: 10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory Psychol.* 10, 639–667. doi: 10.1177/0959354300105004

Michell, J. (2003). The quantitative imperative: positivism, naive realism and the place of qualitative methods in psychology. *Theory Psychol.* 13, 5–31. doi: 10.1177/0959354303013001758

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Front. Psychol.* 3:261. doi: 10.3389/fpsyg.2012.00261

Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: response to Humphry and Sijtsma. *Theory Psychol.* 24, 111–123. doi: 10.1177/0959354313517524

Michell, J. (2020). Thorndike's Credo: metaphysics in psychometrics. *Theory Psychol.* 30, 309–328. doi: 10.1177/0959354320916251

Michell, J. (2022). "The art of imposing measurement upon the mind": Sir Francis Galton and the genesis of the psychometric paradigm. *Theory Psychol.* 32, 375–400. doi: 10.1177/09593543211017671

Michell, J. (2023). "Professor Spearman has drawn over-hasty conclusions": unravelling psychometrics' "Copernican Revolution". *Theory Psychol.* 33, 661–680. doi: 10.1177/09593543231179446

Michell, J., and Ernst, C. (1996). The axioms of quantity and the theory of measurement, Part I. An English translation of holder (1901), Part I. *J. Math. Psychol.* 40, 235–252. doi: 10.1006/jmps.1996.0023

Michell, J., and Ernst, C. (1997). The axioms of quantity and the theory of measurement, Part II. An English translation of holder (1901), Part II. *J. Math. Psychol.* 41, 345–356. doi: 10.1006/jmps.1997.1178

Miller, A. (2024). "Realism", in *The Stanford Encyclopedia of Philosophy*, eds. E. N. Zalta and U. Nodelman. Available online at: https://plato.stanford.edu/archives/sum2024/entries/realism (Accessed May 13, 2025).

Mislevy, R. J. (2009). *Validity from the perspective of model-based reasoning.* CRESST Report 752. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA. Available online at: https://eric.ed.gov/?id=ED507085 (Accessed June 13, 2025).

Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement.* New York/London: Routledge.

Mislevy, R. J. (2019). Advances in the science of measurement and cognition. *Ann. Am. Acad. Politics Soc. Sci.* 683, 164–182. doi: 10.1177/0002716219843816

Mislevy, R. J. (2024). Sociocognitive and argumentation perspectives on psychometric modeling in educational assessment. *Psychometrika* 89, 64–83. doi: 10.1007/s11336-024-09966-5

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Measurement* 2, 201–218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. C. M. (2008). On the implications of the classical ergodic theorems: analysis of developmental processes has to focus on intra-individual variation. *Dev. Psychobiol.* 50, 60–69. doi: 10.1002/dev.20262

Molenaar, P. C. M., and Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Curr. Dir. Psychol. Sci.* 18, 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Montuori, A. (2008). "Foreword," in *Transdisciplinarity. Theory and Practice*, ed. B. Nicolescu (Cresskill, NJ: Hampton Press), IX–XVII.

Moore, S., Speelman, C. P., and McGann, M. (2023). Pervasiveness of effects in sample-based experimental psychology: a re-examination of replication data from nine famous psychology experiments. *New Ideas Psychol.* 68:100978. doi: 10.1016/j.newideapsych.2022.100978

Muthukrishna, M., and Henrich, J. (2019). A problem in theory. *Nat. Hum. Behav.* 3, 221–229. doi: 10.1038/s41562-018-0522-1

Narens, L. (2002). A meaningful justification for the representational theory of measurement. *J. Math. Psychol.* 46, 746–768. doi: 10.1006/jmps.2002.1428

Newell, D., and Tiesinga, E. (2019). *The International System of Units (SI), 2019 Edition, Special Publication (NIST SP).* Gaithersburg, MD: National Institute of Standards and Technology.

Newton, P. E., and Baird, J. A. (2016). The great validity debate. *Assess. Educ. Princ. Policy Pract.* 23, 173–177. doi: 10.1080/0969594X.2016.1172871

Nicolescu, B. (2008). *Transdisciplinarity – Theory and Practice*. Cresskill, NJ: Hampton Press.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science* 348, 1422–1425. doi: 10.1126/science.aab2374

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Pace, V. L., and Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Pers. Individ. Dif.* 49, 669–676. doi: 10.1016/j.paid.2010.06.014

Parsons, C. (1990). The structuralist view of mathematical objects. *Synthese* 84, 303–346 doi: 10.1007/BF00485186

Peirce, C. S. (1958). *Collected Papers of Charles Sanders Peirce*, Vols. 1-6, eds. C. Hartshorne and P. Weiss, Vols. 7-8, ed. A. W. Burks. Cambridge, MA: Harvard University Press.

Piaget, J. (1972). "The epistemology of interdisciplinary relationships," in *Centre for Educational Research and Innovation (CERI). Interdisciplinarity: Problems of Teaching and Research in Universities* (Paris, France: Organisation for Economic Co-operation and Development), 127–139.

Poli, R. (2001). Foreword. *Axiomathes* 12, 1–5. doi: 10.1023/A:1015841116773

Poli, R., and Seibt, J. (Eds.) (2010). *Theory and Applications of Ontology: Philosophical Perspectives*. Dordrecht; Heidelberg; London; New York, NY: Springer Verlag.

Poovey, M. (1998). *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. Chicago: Chicago University Press.

Porter, T. (1995). *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life.* Princeton, N.J: Princeton University Press.

Ramage, M., and Shipp, K. (2020). *Systems Thinkers* (2nd Edn.). London, UK: Springer.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests. Studies Im Mathematical Psychology*. Kopenhagen: Danmarks pædagogiske Institut.

Resnick, M. D. (1997). *Mathematics as a Science of Patterns.* Oxford, England: Clarendon.

Revelle, W. (2024). The seductive beauty of latent variable models: or why I don't believe in the Easter Bunny. *Pers. Indiv. Diff.* 221, 112552, 1–17. doi: 10.1016/j.paid.2024.112552

Richters, J. E. (2021). Incredible utility: the lost causes and causal debris of psychological science. *Basic Appl. Soc. Psych.* 43, 366–405. doi: 10.1080/01973533.2021.1979003

Robinson, O. C. (2011). The idiographic/nomothetic dichotomy: tracing historical origins of contemporary confusions. *History Philos. Psychol.* 13, 32–39. doi: 10.53841/bpshpp.2011.13.2.32

Rogoff, B. (2011). *Developing Destinies: A Mayan Midwife and Town*. Oxford: Oxford University Press.

Rose, T. (2016). *The End of Average: How to Succeed in a World that Values Sameness.* New York: Harper Collins.

Rosen, R. (1985). *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations.* New York: Elsevier Science and Technology Books.

Rosen, R. (1991). *Life Itself. A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life.* New York: Columbia University Press.

Rosen, R. (1999). *Essays on Life Itself*. New York, NY: Columbia University Press.

Rudmin, F., Trimpop, R. M., Kryl, I., and Boski, P. (1987). Gustav Ichhieser in the history of social psychology: an early phenomenology of social attribution. *Br. J. Soc. Psychol.* 26, 165–180. doi: 10.1111/j.2044-8309.1987.tb00777.x

Rudolph, L. (2013). "Qualitative mathematics for the social sciences," in *Mathematical Models for Research on Cultural Dynamics,* ed. L. Rudolph (London: Routledge), 492.

Salvatore, S., and Valsiner, J. (2010). Between the general and the unique: overcoming the nomothetic versus idiographic opposition. *Theory Psychol.* 20, 817–833. doi: 10.1177/0959354310381156

Sato, T., Hidaka, T., and Fukuda, M. (2009). "Depicting the dynamics of living the life: the trajectory equifinality model," in *Dynamic Process Methodology in the Ssocial and Developmental Sciences,* eds. J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, and N. Chaudhary (New York, NY: Springer US), 217–240.

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology* 5:1645. doi: 10.15626/MP.2019.1645

Schönemann, P. H. (1994). "Measurement: the reasonable ineffectiveness of mathematics in the social sciences," in *Trends and Perspectives in Empirical Social Research*, eds. I. Borg and P. P. Mohler (Berlin; New York, NY: De Gruyter), 149–160.

Schrödinger, E. (1964). *What is Life? Reprinted in: What is life? With Mind and Matter and Autobiographical Sketches.* Cambridge, UK: Cambridge University Press.

Schwager, K. W. (1991). The representational theory of measurement: an assessment. *Psychol. Bull.* 110, 618–626. doi: 10.1037/0033-2909.110.3.618

Sechrest, L., McKnight, P., and McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *Am. Psychol.* 51, 1065–1071. doi: 10.1037/0003-066X.51.10.1065

Semin, G. (1989). The contribution of linguistic factors to attribute inference and semantic similarity judgements. *Eur. J. Soc. Psychol.* 19, 85–100. doi: 10.1002/ejsp.2420190202

Serva, M. A., Fuller, M. A., and Mayer, R. C. (2005). The reciprocal nature of trust: a longitudinal study of interacting teams. *J. Organ. Behav.* 26, 625–648. doi: 10.1002/job.331

Sherry, D. (2011). Thermometers and the foundations of measurement. *Philos. Sci.* 78, 512–528.

Shweder, R. A. (1977). Likeness and likelihood in everyday thought: magical thinking in judgments about personality. *Curr. Anthropol.* 18, 637–658. doi: 10.1086/201974

Shweder, R. A., and D'Andrade, R. G. (1980). "The systematic distortion hypothesis," in *Fallible Judgment in Behavioral Research: New Directions for Methodology of Social and Behavioral Science*, ed. R. A. Shweder (San Francisco, CA: Jossey-Bass), 37–58.

Simmel, G. (1900/1978). *The Philosophy of Money (T. Bottomore and D. Frisby, Trans.).* London: Routledge.

Skinner, E. A. (1996). A guide to constructs of control. *J. Pers. Soc. Psychol.* 71, 549–570. doi: 10.1037/0022-3514.71.3.549

Slaney, K. L. (2017). "Some conceptual housecleaning," in *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions* (London: Palgrave Macmillan), 201–234.

Smedslund, G., Arnulf, J. K., and Smedslund, J. (2022). Is psychological science progressing? Explained variance in PsycINFO articles during the period 1956 to 2022. *Front. Psychol.* 13:1089089. doi: 10.3389/fpsyg.2022.1089089

Smedslund, J. (1978). Bandura's theory of self-efficacy: a set of common sense theorems. *Scand. J. Psychol.* 19, 1–14. doi: 10.1111/j.1467-9450.1978.tb00299.x

Smedslund, J. (1988). *Psycho-Logic*. Berlin, Heidelberg: Springer.

Smedslund, J. (1991). The pseudoempirical in psychology and the case for psychologic. *Psychol. Inq.* 2, 325–338. doi: 10.1207/s15327965pli0204_1

Smedslund, J. (2012). The bricoleur model of psychological practice. *Theory Psychol.* 22, 643–657. doi: 10.1177/0959354312441277

Smedslund, J. (2016). Why psychology cannot be an empirical science. *Integr. Psychol. Behav. Sci.* 50, 185–195. doi: 10.1007/s12124-015-9339-x

Smedslund, J. (2021). From statistics to trust: psychology in transition. *New Ideas Psychol.* 61:100848. doi: 10.1016/j.newideapsych.2020.100848

Spearman, C. (1904). General intelligence, objectively determined and measured. *Am. J. Psychol.* 15, 201–293. doi: 10.2307/1412107

Speelman, C. P., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013.00451

Speelman, C. P., and McGann, M. (2020). Statements about the pervasiveness of behaviour require data about the pervasiveness of behaviour. *Front. Psychol.* 11:594675. doi: 10.3389/fpsyg.2020.594675

Speelman, C. P., Parker, L., Rapley, B. J., and McGann, M. (2024). Most psychological researchers assume their samples are ergodic: evidence from a year of articles in three major journals. *Collabra Psychol.* 10:92888. doi: 10.1525/collabra.92888

Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach.* Oxford: Blackwell.

Steinmetz, G. (ed.) (2005). *The Politics of Method in the Human Sciences: Positivism and its Epistemological Others.* Durham, NC: Duke University Press.

Stern, W. (1911). *Die Differentielle Psychologie in Ihren Methodischen Grundlagen [Differential Psychology in its Methodological Foundations]*. Leipzig: Barth.

Stevens, S. S. (1935). The operational definition of psychological concepts. *Psychol. Rev.* 42, 517–527. doi: 10.1037/h0056973

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Stevens, S. S. (1958). Measurement and man. *Sci. New Series* 127, 383–389. doi: 10.1126/science.127.3295.383

Storozuk, A., Ashley, M., Delage, V., and Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *Quant. Methods Psychol.* 16, 472–481. doi: 10.20982/tqmp.16.5.p472

Straub, D. W., Boudreau, M.-C., and Gefen, D. (2004). Validation guidelines for IS positivist research. *Commun. Assoc. Inf. Syst.* 13, 380–427. doi: 10.17705/1CAIS.01324

Sugarman, J. (2017). Psychologism as a style of reasoning and the study of persons. *New Ideas Psychol.* 44, 21–27. doi: 10.1016/j.newideapsych.2016.11.008

Sugarman, J., and Martin, J. (2020). *A Humanities Approach to the Psychology of Personhood*. London: Routledge.

Suppes, P., Krantz, D., Luce, D., and Tversky, A. (1989). *Foundations of Measurement, Vol. II: Geometrical, Threshold, and Probabilistic Representations*. New York, NY: Academic Press.

Suppes, P., and Zinnes, J. (1963). "Basic measurement theory," in *Handbook of Mathematical Psychology*, ed. D. Luce (New York, NY: John Wiley and Sons), 1–76.

Tal, E. (2020). "Measurement in science," in *The Stanford Encyclopedia of Philosophy (Fall 2020). Metaphysics Research Lab, Stanford University,* ed. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/fall2020/entries/measurement-science (Accessed October 21, 2023).

Tang, R., and Braver, T. S. (2020). Towards an individual differences perspective in mindfulness training research: theoretical and empirical considerations. *Front. Psychol.* 11:818. doi: 10.3389/fpsyg.2020.00818

Taylor, C. (1985). "Peaceful coexistence in psychology," in *Human Agency and Language*, ed. C. Taylor (New York: Cambridge University Press), 117–138.

Thomas, M. A. (2019). Mathematization, not measurement: a critique of Stevens' scales of measurement. *J. Methods Meas. Soc. Sci.* 10, 76–94. doi: 10.2458/v10i2.23785

Thorndike, E. L. (1903). *Notes on Child Study* (2nd Edn.). New York, NY: Macmillan.

Thurstone, L. L. (1928). Attitudes can be measured. *Am. J. Sociol.* 33, 529–554. doi: 10.1086/214483

Tolman, C. W. (ed.) (1992). *Positivism in Psychology: Historical and Contemporary Problems.* New York, NY: Springer.

Toomela, A. (2000). Activity theory is a dead end for cultural-historical psychology. *Cult. Psychol.* 6, 353–364. doi: 10.1177/1354067X0063005

Toomela, A. (2003). "Development of symbol meaning and the emergence of the semiotically mediated mind," in *Cultural Guidance in the Development of the Human Mind,* ed. A. Toomela (Westport, CT: Ablex Publishing), 163–209.

Toomela, A. (2007a). Culture of science: strange history of the methodological thinking in psychology. *Integr. Psychol. Behav. Sci.* 41, 6–20. doi: 10.1007/s12124-007-9004-0

Toomela, A. (2007b). Sometimes one is more than two: when collaboration inhibits knowledge construction. *Integr. Psychol. Behav. Sci.* 41, 198–207. doi: 10.1007/s12124-007-9015-x

Toomela, A. (2007c). "Unifying psychology: absolutely necessary, not only useful," in *Psicologia: Novas direcoes no dialogo com outros campos de saber*, eds. A. V. B. Bastos and N. M. D. Rocha (São Paulo: Casa do Psicologo), 449–464.

Toomela, A. (2008a). Activity theory is a dead end for methodological thinking in cultural psychology too. *Cult. Psychol.* 14, 289–303. doi: 10.1177/1354067X08088558

Toomela, A. (2008b). Variables in psychology: a critique of quantitative psychology. *Integr. Psychol. Behav. Sci.* 42, 245–265. doi: 10.1007/s12124-008-9059-6

Toomela, A. (2008c). Vygotskian cultural-historical and sociocultural approaches represent two levels of analysis: complementarity instead of opposition. *Culture and Psychology* 14, 57–69. doi: 10.1177/1354067X07085812

Toomela, A. (2009a). "How methodology became a toolbox - and how it escapes from that box," in *Dynamic Process Methodology in the Social and Developmental Sciences*, eds. J. Valsiner, P. Molenaar, M. Lyra, and N. Chaudhary (New York: Springer), 45–66.

Toomela, A. (2009b). "Kurt Lewin's contribution to the methodology of psychology: from past to future skipping the present," in *The Observation of Human Systems. Lessons from the History of Anti-Reductionistic Empirical Psychology*, ed. J. Clegg (New Brunswick, NJ: Transaction Publishers), 101–116.

Toomela, A. (2010). "Modern mainstream psychology is the best? Noncumulative, historically blind, fragmented, atheoretical," in *Methodological Thinking in Psychology: 60 Years Gone Astray?,* eds. A. Toomela and J. Valsiner (Charlotte: Information Age Publishers), 1–26.

Toomela, A. (2011). Travel into a fairy land: a critique of modern qualitative and mixed methods psychologies. *Integr. Psychol. Behav. Sci.* 45, 21–47. doi: 10.1007/s12124-010-9152-5

Toomela, A. (2012). "Guesses on the future of cultural psychology: past, present, and past," in *The Oxford Handbook of Culture and Psychology*, ed. J. Valsiner (New York: Oxford University Press), 998–1033.

Toomela, A. (2014a). "Mainstream psychology," in *Encyclopedia of Critical Psychology*, ed. T. Teo (New York: Springer), 1117–1125.

Toomela, A. (2014b). Methodology of cultural-historical psychology. In A. Yasnitsky, R. van der Veer, and M. Ferrari (Eds.), *The Cambridge Handbook of Cultural-Historical Psychology* (pp. 99-125). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139028097.007

Toomela, A. (2014c). "Modern qualitative approach to psychology: art or science?," in *Multicentric Identities in a Globalizing World*, eds. S. Salvatore, A. Gennaro, and J. Valsiner (Charlotte, NC: Information Age Publishing), 75–82.

Toomela, A. (2014d). "A structural systemic theory of causality and catalysis," in *The Catalyzing Mind. Beyond Models of Causality,* eds. K. R. Cabell and J. Valsiner (New York: Springer), 271–292.

Toomela, A. (2015). Vygotsky's theory on the Procrustes' bed of linear thinking: looking for structural-systemic theseus to save the idea of 'social formation of mind'. *Cult. Psychol.* 21, 318–339. doi: 10.1177/1354067X15570490

Toomela, A. (2016a). "The ways of scientific anticipation: from guesses to probabilities and from there to certainty," in *Anticipation Across Disciplines,* ed. M. Nadin (Cham: Springer. 255–273.

Toomela, A. (2016b). What are higher psychological functions? *Integr. Psychol. Behav. Sci.* 50, 91–121. doi: 10.1007/s12124-015-9328-0

Toomela, A. (2017). "Towards general-unifying theory of psychology: Engelsted and beyond," in *Catching up with Aristotle. A Journey in Quest for General Psychology,* ed. N. Engelsted (Cham: Springer), 137–150.

Toomela, A. (2019). *The Psychology of Scientific Inquiry.* Cham: Springer Nature.

Toomela, A. (2020). *Culture, Speech and My Self*. Sepamäe: Porcos ante Margaritas.

Toomela, A. (2022). "Methodology of science: different kinds of questions require different methods," in *Experimental Psychology: Ambitions and Possibilities,* eds. D. Gozli and J. Valsiner (Chum: Springer), 113–151.

Toomela, A., and Valisiner, J. (2010). Methodological thinking in psychology: 60 years gone astray? US: IAP.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York, NY: Wiley.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theor. Psychol.* 19, 579–599. doi: 10.1177/0959354309341926

Trendler, G. (2013). Measurement in psychology: a case of ignoramus et ignorabimus? A rejoinder. *Theor. Psychol.* 23, 591–615. doi: 10.1177/0959354313490451

Trendler, G. (2019a). Conjoint measurement undone. *Theor. Psychol.* 29, 100–128. doi: 10.1177/0959354318788729

Trendler, G. (2019b). Measurability, systematic error, and the replication crisis: a reply to Michell (2019) and Krantz and Wallsten (2019). *Theor. Psychol.* 29, 144–151. doi: 10.1177/0959354318824414

Trendler, G. (2022a). Is measurement in psychology an empirical or a conceptual issue? A comment on David Franz. *Theor. Psychol.* 32, 164–170. doi: 10.1177/09593543211050025

Trendler, G. (2022b). The incoherence of Rasch measurement: a critical comparison between measurement in psychology and physics. *Pers. Individ. Dif.* 189:111408. doi: 10.1016/j.paid.2021.111408

Uher, J. (2011). Individual behavioral phenotypes: an integrative meta-theoretical framework. Why 'behavioral syndromes' are not analogues of 'personality'. *Dev. Psychobiol.* 53, 521–548. doi: 10.1002/dev.20544

Uher, J. (2013). Personality psychology: lexical approaches, assessment methods, and trait concepts reveal only half of the story-Why it is time for a paradigm shift. *Integr. Psychol. Behav. Sci.* 47, 1–55. doi: 10.1007/s12124-013-9230-6

Uher, J. (2015a). Conceiving "personality": psychologist's challenges and basic fundamentals of the transdisciplinary philosophy-of-science paradigm for research on individuals. *Integr. Psychol. Behav. Sci.* 49, 398–458. doi: 10.1007/s12124-014-9283-1

Uher, J. (2015b). Developing "personality" taxonomies: metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integr. Psychol. Behav. Sci.* 49, 531–589. doi: 10.1007/s12124-014-9280-4

Uher, J. (2015c). Interpreting "personality" taxonomies: why previous models cannot capture individual-specific experiencing, behaviour, functioning and development. Major taxonomic tasks still lay ahead.. *Integr. Psychol. Behav. Sci.* 49, 600–655. doi: 10.1007/s12124-014-9281-3

Uher, J. (2015d). "Agency enabled by the psyche: explorations using the transdisciplinary philosophy-of-science paradigm for research on individuals," in *Constraints of Agency: Explorations of Theory in Everyday Life. Annals of Theoretical Psychology,* Vol. 12, eds. C. W. Gruber, M. G. Clark, S. H. Klempe, and J. Valsiner (New York: Springer International Publishing), 177–228.

Uher, J. (2016a). "Exploring the workings of the Psyche: metatheoretical and methodological foundations," in *Psychology as the Science of Human Being: The Yokohama Manifesto*, eds. J. Valsiner, G. Marsico, N. Chaudhary, T. Sato, and V. Dazzani (New York: Springer International Publishing), 299–324.

Uher, J. (2016b). What is behaviour? And (when) is language behaviour? A metatheoretical definition. *J. Theory Soc. Behav.* 46, 475–501. doi: 10.1111/jtsb.12104

Uher, J. (2018a). Quantitative data from rating scales: an epistemological and methodological enquiry. *Front. Psychol.* 9:2599. doi: 10.3389/fpsyg.2018.02599

Uher, J. (2018b). Taxonomic models of individual differences: a guide to transdisciplinary approaches. *Philos. Trans. Royal Soc. B* 373:20170171. doi: 10.1098/rstb.2017.0171

Uher, J. (2018c). "The transdisciplinary philosophy-of-science paradigm for research on individuals: foundations for the science of personality and individual differences," in *The SAGE Handbook of Personality and Individual Differences: Volume I: The science of Personality and Individual Differences,* eds. V. Zeigler-Hill and T. K. Shackelford (London, UK: SAGE), 84–109.

Uher, J. (2019). Data generation methods across the empirical sciences: differences in the study phenomena's accessibility and the processes of data encoding. *Qual. Quan. Int. J. Methodol.* 53, 221–246. doi: 10.1007/s11135-018-0744-3

Uher, J. (2020a). Measurement in metrology, psychology and social sciences: data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Qual. Quan.* 54, 975–1004. doi: 10.1007/s11135-020-00970-2

Uher, J. (2020b). Human uniqueness explored from the uniquely human perspective: epistemological and methodological challenges. *J. Theory Soc. Behav.* 50, 20–24. doi: 10.1111/jtsb.12232

Uher, J. (2021a). Psychometrics is not measurement: unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.* 41:58. doi: 10.1037/teo0000176

Uher, J. (2021b). Quantitative psychology under scrutiny: measurement requires not result-dependent but traceable data generation. *Pers. Individ. Dif.* 170:110205. doi: 10.1016/j.paid.2020.110205

Uher, J. (2021c). Psychology's status as a science: peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integr. Psychol. Behav. Sci.* 55, 212–224. doi: 10.1007/s12124-020-09545-0

Uher, J. (2021d). Problematic research practices in psychology: misconceptions about data collection entail serious fallacies in data analyses. *Theor. Psychol.* 31, 411–416. doi: 10.1177/09593543211014963

Uher, J. (2022a). Functions of units, scales and quantitative data: fundamental differences in numerical traceability between sciences. *Qual. Quan.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: a rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

Uher, J. (2023a). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Soc. Personal. Psychol. Compass* 17:e12740. doi: 10.1111/spc3.12740

Uher, J. (2023b). What are constructs? Ontological nature, epistemological challenges, theoretical foundations and key sources of misunderstandings and confusions. *Psychol. Inq.* 34, 280–290. doi: 10.1080/1047840X.2023.2274384

Uher, J. (2024). "Transdisciplinarity, complexity thinking and dialectics," in *The Routledge International Handbook of Dialectical Thinking*, eds. N. Shannon, M. Mascolo, and A. Belolutskaya (London: Routledge), 259–277. doi: 10.4324/9781003317340-21

Uher, J. (2025). Statistics is not measurement: the inbuilt semantics of psychometric scales and language-based models obscures crucial epistemic differences. *Front. Psychol.* 16:1534270. doi: 10.3389/fpsyg.2025.1534270

Valsiner, J. (1998). *The Guide Mind. A Sociogenetic Approach to Personality*. Cambridge, MA: Harvard University Press.

Valsiner, J. (2000). *Culture and Human Development*. London: Sage.

Valsiner, J. (2007). *Culture in Minds and Societies. Foundations of Cultural Psychology*. Thousand Oaks, CA: Sage.

Valsiner, J. (2012). *A Guided Science: History of Psychology in the Mirror of Its Making*. New Brunswick, NJ: Transaction Publishers.

Valsiner, J. (2014a). *An Invitation to Cultural Psychology*. London: SAGE Publications.

Valsiner, J. (2014b). Needed for cultural psychology: methodology in a new key. *Cult. Psychol.* 20, 3–30. doi: 10.1177/1354067X13515941

Valsiner, J. (2017). *From Methodology to Methods in Human Psychology*. Cham: Springer.

Valsiner, J., and Brinkmann, S. (2016). "Beyond the "variables": developing metalanguage for psychology," in *Centrality of History for Theory Construction in Psychology, Annals of Theoretical Psychology*, eds. S. Klempe and R. Smith (New York: Springer), 75–90.

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective*. New York: Oxford University Press.

van Geert, P. (2011). The contribution of complex dynamic systems to development. *Child Dev. Perspect.* 5, 273–278. doi: 10.1111/j.1750-8606.2011.00197.x

van Inwagen, P., Sullivan, M., and Bernstein, S. (2023). "Metaphysics," in *The Stanford Encyclopedia of Philosophy (Summer 2023 Edition)*, eds. E. N. Zalta and U. Nodelman. Available online at: https://plato.stanford.edu/archives/sum2023/entries/metaphysics (Accessed May 18, 2025).

Velleman, P. F., and Wilkinson, L. (1993). Nominal ordinal interval and ratio typologies are misleading. *Am. Stat.* 47, 65–72. doi: 10.1080/00031305.1993.10475938

Vessonen, E. (2017). Psychometrics versus representational theory of measurement. *Philos. Soc. Sci.* 47, 330–350. doi: 10.1177/0048393117705299

von Eye, A., and Bergman, L. R. (2003). Research strategies in developmental psychopathology: dimensional identity and the person-oriented approach. *Dev. Psychopathol.* 15, 553–580. doi: 10.1017/S0954579403000294

von Eye, A., and Bogat, G. A. (2006). Person-oriented and variable-oriented research: concepts, results, and development. *Merrill-Palmer Q.* 52, 390–420. doi: 10.1353/mpq.2006.0032

von Helmholtz, H. (1887). "Zählen und Messen, erkenntnisstheoretisch betrachtet," in *Philosophische Aufsätze, Eduard Zeller zu seinem fünfzigjährigen Doctorjubiläum gewidmet,* ed. F. T. von Vischer (Leipzig: Fues' Verlag), 17–52.

von Kries, J. (1882). Ueber die Messung intensiver Grössen und über das sogenannte psychophysische Gesetz. *Viertelj. Wiss. Philos.* 6, 256–294.

von Neumann, J. (1955). *Mathematical foundations of quantum mechanics [original: Mathematische Grundlagen der Quantenmechanik in 1935]*. Princeton, NJ: Princeton University Press.

Vygotsky, L. S. (1962). *Thought and Language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1982). "Istoricheski smysl psikhologicheskogo krizisa. Metodologicheskoje issledovanije. (Historical meaning of the crisis in psychology. A methodological study. Originally written in 1927; First published in 1982," in *L. S. Vygotsky. Sobranije sochinenii. Tom 1. Voprosy teorii i istorii psikhologii*, eds. A. R. Luria and M. G. Jaroshevskii (Moscow: Pedagogika), 291–436.

Vygotsky, L. S. (1994). "The problem of the cultural development of the child (Originally published in 1929)," in *The Vygotsky Reader,* eds. R. v. d. Veer and J. Valsiner (Oxford: Blackwell), 57–72.

Vygotsky, L. S. (1997). "The historical meaning of the crisis in psychology: a methodological investigation," in *The Collected Works of L. S. Vygotsky. Volume 3. Problems of the Theory and History of Psychology,* eds. R. W. Rieber and J. Wollock (New York: Springer), 233–343.

Weber, M. (1904–05/1992). *The Protestant Ethic and the Spirit of Capitalism (T. Parsons, Trans.)*. New York: Routledge.

Weber, M. (1949). *The Methodology of the Social Sciences [Translated and edited by E.A. Shils and H.A. Finch]*. New York: Free Press.

Weber, R. (2012). Evaluating and developing theories in the information systems discipline. *J. Assoc. Inf. Syst.* 13, 1–30. doi: 10.17705/1jais.00284

Weber, R. (2021). Constructs and indicators: an ontological analysis. *MIS Q.* 45, 1644–1678. doi: 10.25300/MISQ/2021/15999

Werner, H. (1948). *Comparative Psychology of Mental Development.* New York: International Universities Press.

White, R. (2011). The meaning of measurement in metrology. *Accred. Qual. Assur.* 16, 31–41. doi: 10.1007/s00769-010-0698-1

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. New York, NY: Oxford University Press.

Williams, R. N. (2015). "Introduction," in *Scientism: The New Orthodoxy,* eds. R. N. Williams and D. N. Robinson (New York, NY: Bloomsbury Academic), 1–21.

Windelband, W. (1904/1998). History and Natural Science. *Theor. Psychol.* 8, 5–22. doi: 10.1177/0959354398081001

Wittgenstein, L. (1953). *Philosophical Investigations (G. E. M. Anscombe Trans.).* New York, NY: Wiley-Blackwell.

Wundt, W. (1897). *Outlines of Psychology*. Leipzig: Wilhelm Engelman.

Yarkoni, T. (2022). The generalizability crisis. *Behav. Brain Sci.* 45:e1. doi: 10.1017/S0140525X20001685

Zell, E., Krizan, Z., and Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *Am. Psychol.* 70, 10–20. doi: 10.1037/a0038208

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2017). Making replication mainstream. *Behav. Brain Sci.* 1–50. doi: 10.31234/osf.io/4tg9c

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., et al. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychon. Bull. Rev.* 25, 1968–1972. doi: 10.3758/s13423-017-1348-y

# Frontiers in Psychology

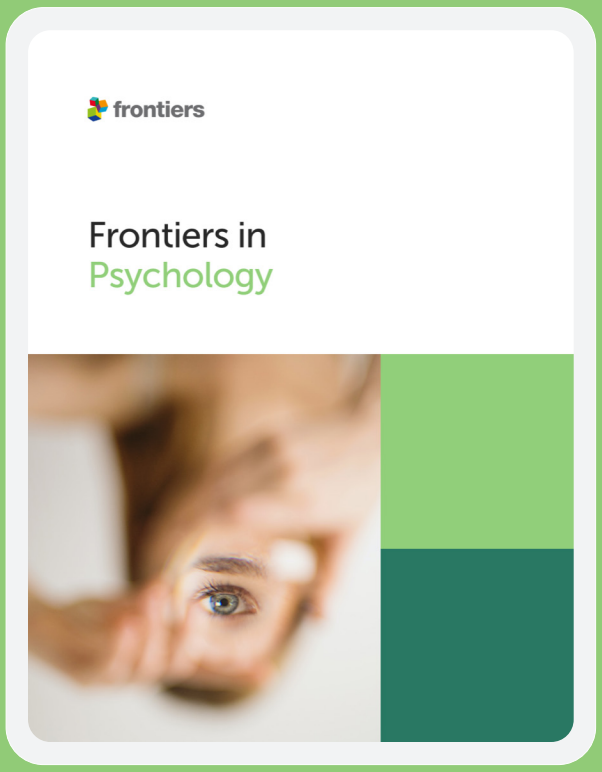
**Paving the way for a greater understanding of human behavior**

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

---

## Discover the latest Research Topics

See more →




**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact