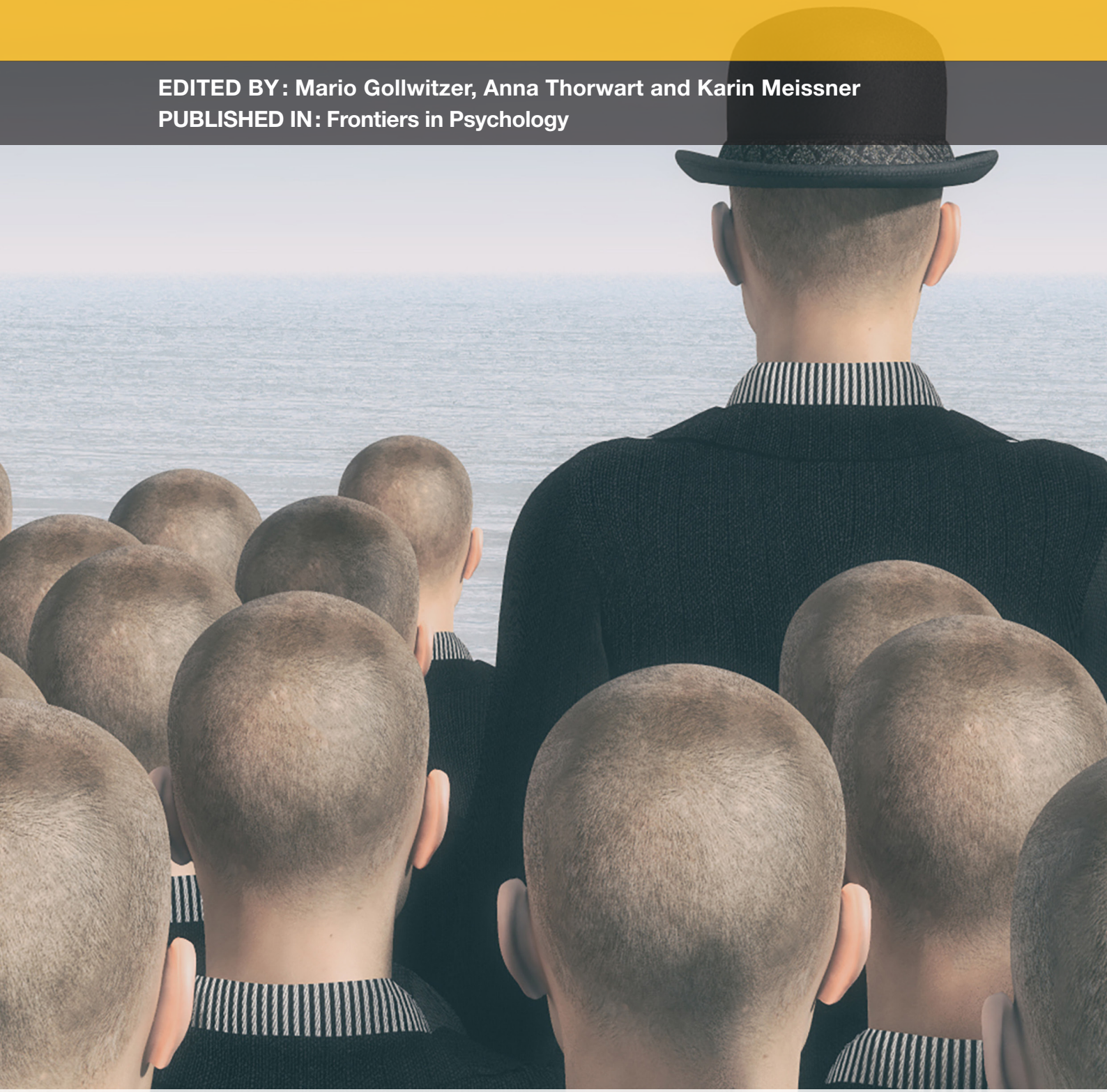
PSYCHOLOGICAL RESPONSES TO VIOLATIONS OF EXPECTATIONS: PERSPECTIVES AND ANSWERS FROM DIVERSE FIELDS OF PSYCHOLOGY

EDITED BY: Mario Gollwitzer, Anna Thorwart and Karin Meissner
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics



frontiers

Frontiers Copyright Statement

© Copyright 2007-2018 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714
ISBN 978-2-88945-445-7
DOI 10.3389/978-2-88945-445-7

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: **researchtopics@frontiersin.org**

PSYCHOLOGICAL RESPONSES TO VIOLATIONS OF EXPECTATIONS: PERSPECTIVES AND ANSWERS FROM DIVERSE FIELDS OF PSYCHOLOGY

Topic Editors:
Mario Gollwitzer, Philipps University of Marburg, Germany
Anna Thorwart, Philipps University of Marburg, Germany
Karin Meissner, LMU Munich, University of Applied Sciences Coburg, Germany



Image: Orla/Shutterstock.com

From Pavlov's dog expecting food when hearing a bell to stereotypes as expectations about other people's behaviour, from Bandura's self-efficacy as expectation for success and failure of one's own behaviour to the "predictive brain" concept in current perception theories: expectations have been a central construct in different areas of psychological research. In each of these areas, specific concepts, theoretical approaches, and empirical methods have been developed to explain when and why expectations persist and when they do not.

Many theories assume that expectations are likely to change in the face of disconfirming evidence. However, sometimes expectations persist even though they are empirically violated, suggesting that they can be "sticky" under certain circumstances. But what are these circumstances? And what are the psychological mechanisms that can explain why and when expectations persist or change after being confronted with expectation-violating evidence?

Each contribution of the current book offers insights into individuals' reactions to violations of expectations. They show that many pieces of the puzzle have been collected in the many sub-displiclines of psychology and that putting them together in an integrative fashion stays a fascinating enterprise.

Citation: Gollwitzer, M., Thorwart, A., Meissner, K., eds. (2018). Psychological Responses to Violations of Expectations: Perspectives and Answers from Diverse Fields of Psychology. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-445-7

Table of Contents

Chapter I: Editorial

06 ***Editorial: Psychological Responses to Violations of Expectations***
Mario Gollwitzer, Anna Thorwart and Karin Meissner

Chapter II: Clinical Psychology

09 ***On the Maintenance of Expectations in Major Depression – Investigating a Neglected Phenomenon***
Tobias Kube, Winfried Rief and Julia A. Glombiewski

16 ***Can Psychological Expectation Models Be Adapted for Placebo Research?***
Winfried Rief and Keith J. Petrie

22 ***Placebo and Nocebo Effects: The Advantage of Measuring Expectations and Psychological Factors***
Nicole Corsi and Luana Colloca

33 ***Patients' Expectations Regarding Medical Treatment: A Critical Review of Concepts and Their Assessment***
Johannes A. C. Laferton, Tobias Kube, Stefan Salzmann, Charlotte J. Auer and Meike C. Shedden-Mora

Chapter III: Social and Personality Psychology

45 ***Expectations of Social Inclusion and Exclusion***
Eric D. Wesselmann, James H. Wirth and Michael J. Bernstein

50 ***Expectations and Decisions in the Volunteer's Dilemma: Effects of Social Distance and Social Projection***
Joachim I. Krueger, Johannes Ullrich and Leonard J. Chen

64 ***Cognitive Load Does Not Affect the Behavioral and Cognitive Foundations of Social Cooperation***
Laura Mieth, Raoul Bell and Axel Buchner

78 ***Trustworthy Tricksters: Violating a Negative Social Expectation Affects Source Memory and Person Perception When Fear of Exploitation Is High***
Philipp Süssenbach, Mario Gollwitzer, Laura Mieth, Axel Buchner and Raoul Bell

90 ***Functional Significance of Conflicting Age and Wealth Cross-Categorization: The Dominant Role of Categories That Violate Stereotypical Expectations***
Jingjing Song and Bin Zuo

105 ***How Do Discrepancies between Victimization and Rejection Expectations in Gay and Bisexual Men Relate to Mental Health Problems?***
Frank A. Sattler and Hanna Christiansen

111 ***Expectation Violation in Political Decision Making: A Psychological Case Study***
Michael Öllinger, Karin Meissner, Albrecht von Müller and Carlos Collado Seidel

Chapter IV: Cognitive and Experimental Psychology

120 ***Three Ways That Non-associative Knowledge May Affect Associative Learning Processes***
Anna Thorwart and Evan J. Livesey

135 ***Self-Generated or Cue-Induced—Different Kinds of Expectations to Be Considered***
Maike Kemper and Robert Gaschler

139 ***Reminder Cues Modulate the Renewal Effect in Human Predictive Learning***
Javier Bustamante, Metin Uengoer and Harald Lachnit

148 ***Compound Stimulus Presentation Does Not Deepen Extinction in Human Causal Learning***
Oren Griffiths, Nathan Holmes and R. Fred Westbrook

161 ***Learning to Detect Triggers of Airway Symptoms: The Role of Illness Beliefs, Conceptual Categories and Actual Experience with Allergic Symptoms***
Thomas Janssens, Eva Caris, Ilse Van Diest and Omer Van den Bergh

171 ***How You Move Is What I See: Planning an Action Biases a Partner's Visual Search***
Dominik Dötsch, Cordula Vesper and Anna Schubö

183 ***Task-Irrelevant Expectation Violations in Sequential Manual Actions: Evidence for a "Check-after-Surprise" Mode of Visual Attention and Eye-Hand Decoupling***
Rebecca M. Foerster

Chapter V: Neurosciences

195 ***Violations of Expectations As Matter for the Believing Process***
Hans-Ferdinand Angel and Rüdiger J. Seitz

207 ***Learning about Expectation Violation from Prediction Error Paradigms – A Meta-Analysis on Brain Processes Following a Prediction Error***
Lisa D'Astolfo and Winfried Rief

218 ***The Dopamine Prediction Error: Contributions to Associative Models of Reward Learning***
Helen M. Nasser, Donna J. Calu, Geoffrey Schoenbaum and Melissa J. Sharpe

frontiers in Psychology

EDITORIAL
published: 23 January 2018
doi: 10.3389/fpsyg.2017.02357

Editorial: Psychological Responses to Violations of Expectations

Mario Gollwitzer[1]*, *Anna Thorwart*[1] and *Karin Meissner*[2,3]

[1] Department of Psychology, Philipps University of Marburg, Marburg, Germany, [2] Institute of Medical Psychology, Ludwig-Maximilians-Universität München, Munich, Coburg, Germany, [3] Division of Health Promotion, Department of Social Work and Health, Hochschule Coburg, Coburg, Germany

Keywords: expectation violation, associative learning, clinical psychology, social psychology, individual differences

Editorial on the Research Topic

Psychological Responses to Violations of Expectations

The general aim of this Research Topic was to collect and systematize theoretical approaches and latest empirical evidence on expectation violations, or, more precisely, on how individuals cope with such violations. This question is relevant from a basic science as well as from an applied perspective. Sometimes, expectations persist even in the face of disconfirming evidence. For instance, social stereotypes remain sticky even after confronting stereotype-inconsistent exemplars, and fear-related expectations are hard to tackle in the course of psychotherapeutic interventions. What are the psychological mechanisms underlying a sustainable change of expectations vs. a persistence of expectations in the face of disconfirming evidence?

The 21 articles collected in the present Research Topic shed more light on this question. As guest editors of this Topic, we were glad to receive papers from so many sub-disciplines of psychology, including clinical psychology (Corsi and Colloca; Kube et al.; Rief and Petrie; Laferton et al.), social/personality psychology (Krueger et al.; Mieth et al.; Song and Zuo; Süssenbach et al.; Wesselmann et al.), learning psychology (Bustamante et al.; Griffiths et al.; Janssens et al.; Kemper and Gaschler; Thorwart and Livesey), cognitive psychology (Dötsch et al.; Foerster), and neurosciences (Angel and Seitz; D'Astolfo and Rief; Nasser et al.), and one paper even builds a bridge to political science (Öllinger et al.).

These papers also cover a broad range of methodological approaches, from theoretical discussion (e.g., Öllinger et al.; Angel and Seitz) via highly controlled lab studies (e.g., Foerster) and surveys (e.g., Sattler and Christiansen) to meta-analyses (e.g., D'Astolfo and Rief). The diversity of specific research questions, theoretical approaches, and methodological strategies is enormous and shows how prevalent expectation violations are and how relevant a psychological model for people's psychological responses to these expectations actually is.

That said, a common theoretical framework on how individuals process and deal with expectation violations is missing. Such a framework would be helpful to (1) establish a common language with properly defined concepts that can be usefully applied to psychological research on expectation violations in different areas, (2) describe the cognitive, affective, and social processes involved in individuals' responses to expectation violations, and (3) explain these responses psychologically. Such a model should not only be applicable to neuroscientific, but also to cognitive and social psychological approaches.

One model that we think may be helpful in that regard is the ViolEx Model (Rief et al., 2015). The ViolEx model defines expectations as conditional predictions about future events (or "if-*X*-then-*Y*" hypotheses) that may be changed or maintained in the face of disconfirming evidence (i.e., if an event or stimulus *X* is followed by a non-expected outcome $\overline{Y}$). The model differentiates between generalized expectations (e.g., "Whenever other people ask me for help, their intention is

OPEN ACCESS

Edited and reviewed by:
Bernhard Hommel,
Leiden University, Netherlands

***Correspondence:**
Mario Gollwitzer
mario.gollwitzer@uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *28 November 2017*
Accepted: *26 December 2017*
Published: *23 January 2018*

Citation:
Gollwitzer M, Thorwart A and Meissner K (2018) Editorial: Psychological Responses to Violations of Expectations. Front. Psychol. 8:2357. doi: 10.3389/fpsyg.2017.02357

to exploit me") and situation-specific, conditional predictions (e.g., "If I lend this book to my neighbor, he will never bring it back"). In general, only situation-specific predictions (but not generalized expectations) can be directly falsified empirically. If a specific prediction turns out to be correct and the expected outcome occurs, the model predicts that one's generalized expectation is reinforced or stabilized. Expectation violations, on the other hand, do not necessarily result in a change of one's generalized expectation.

Whether expectation change or rather expectation maintenance occurs in a given situation depends on the specific psychological process that is operating. The ViolEx model specifies three of these "coping" processes: accommodation, assimilation, and immunization[1]. Technically speaking, these processes mediate the effect of expectation violations on expectation change vs. maintenance.

Accommodation refers to mechanisms by which individuals adjust their expectation so that it fits to the (unexpected) outcome. Thus, accommodation is the process that underlies expectation change in the context of expectation-inconsistent outcomes and corresponds to what is generally referred to as learning (Thorwart and Livesey).

Assimilation refers to mechanisms by which individuals actively remove any future discrepancies between their expectations and expectation-inconsistent outcomes. This strategy includes (a) avoiding expectation-inconsistent outcomes (e.g., "fear avoidance" in clinical psychology; cf. Vlaeyen and Linton, 2012), and/or (b) actively contributing to a higher likelihood of expectation-consistent outcomes (i.e., "self-fulfilling prophecies;" Stinson et al., 2011; Hechler et al., 2016). Thus, individuals create situations that confirm their current expectations and reduce the effect of an expectation violation.

Immunization refers to mechanisms by which individuals minimize the potential impact of discrepant information on their expectations in a given situation. In the case of "data-oriented immunization," individuals devalue discrepant information (e.g., denying the data or doubting its validity). In the case of "concept-oriented immunization," individuals reframe the conceptual meaning of their expectation so that former discrepant information is no longer diagnostically valid (cf. Greve and Wentura, 2010). For instance, studies from social psychology show that confronting people with stereotype-inconsistent out-group exemplars does not necessarily change their stereotypes; stereotype-inconsistent exemplars are often "subtyped" as atypical exemplars of their respective group (Yzerbyt and Carnaghi, 2007). Thus, subtyping is a form of immunization. Possible implications of such immunization processes are far-reaching and may even comprise misguided political decision making (Öllinger et al.).

Taken together, the ViolEx model assumes that organisms can react to expectation violations by following one of three routes (i.e., accommodation, assimilation, immunization), and only one of these routes (i.e., accommodation) actually leads to a sustainable change in existing expectations.

The ViolEx model further predicts that (a) direct experiences, (b) social (and cultural) influences, and (c) individual differences influence which route an organism "chooses" to follow. In other words, each of these three factors influences the probability with which accommodation, assimilation, or immunization occurs. Technically speaking, these factors moderate the effect of expectation violations on expectation maintenance vs. change.

Direct experiences include current situational expositions or prior experiences with X and/or Y and other stimuli. For example, Griffiths et al. explore whether creating a strong expectation by presenting two separate predictive events simultaneously (X_1 and X_2) results in more accommodation and Bustamante et al. investigate the modulatory impact of different "reminder cues" during expectation-consistent and expectation-inconsistent situations on processing these situations. Other findings show that expectations are changed more rapidly when there were only few expectation violations experienced before (e.g., Thorwart et al., 2017). A factor that it also relevant in this regard (and which is has not been explicitly incorporated into the ViolEx model) is how an initial expectation has been generated in the first place. As Kemper and Gaschler argue, self-generated expectations may be more resistant to change than cue-induced expectations. In line with this argument, Janssens et al. show how pre-existing conceptual beliefs shape expectations generated by a cue, and Thorwart and Livesey offer three solutions for how influences of such information can be incorporated into existing learning models.

Social influences include peers, significant others, the media, or any other social or cultural factors. They are particularly relevant in cases of *social* expectations; for instance, expectations about being socially included (Wesselmann et al.), about others' actions in a social dilemma (Krueger et al.), or about other people's trustworthiness (Mieth et al.; Süssenbach et al.). Using the latter as an example, Krueger et al. show that social distance to others is (negatively) correlated with people's expectations that they will cooperate. Finally, the strength of culturally shared stereotypes strongly predicts the stickiness of expectations (Song and Zuo).

Individual differences include personality traits as well as biological/genetic factors. For instance, victim sensitivity—individuals' disposition to react toward injustice at one's own disadvantage (Schmitt et al., 2005)—is associated with a latent ("generalized") expectation of other people being selfish and untrustworthy (Gollwitzer et al., 2013). As Süssenbach et al. show, victim-sensitive individuals have a better source memory for events in which this latent expectation has been violated. Regarding biological/genetic factors, research shows that personality traits that are related to genetic differences in dopaminergic and serotonergic processes may be critical for inter-individual differences in processing reward-prediction errors (e.g., Müller et al., 2014), which is also true for dopamine- and extraversion-related gene-variants (Müller et al., 2011).

The ViolEx model is a useful framework for different approaches to investigations of expectation violations. This does not mean that it is the best of all possible models. In fact, there

[1]These terns are borrowed from research on coping with age-related stressors (Brandtstädter and Greve, 1994; Rothermund and Brandtstädter, 2003). Since this research does not talk so much about the change vs. persistence in expectations (but rather about the change vs. persistence of goals, plans, and self-concepts), the terms have a slightly different meaning in the present context.

are other models (such as the Credition model portrayed by Angel and Seitz), and, of course, the ViolEx model may need to be adapted to represent the specific aspects of a particular research area. In this vein, Rief and Petrie show how the ViolEx model can be adapted to research on Placebo/Nocebo effects in clinical psychology. Furthermore, the ViolEx model is currently silent on the neuropsychological implementation of its variables and processes as well as its links to other relevant research, for example, on the dopamine prediction error (Nasser et al.).

This Research Topic shows that for scholars in different psychological research areas, investigating individuals' reactions to violations of expectations is a fascinating endeavor. Many pieces of the puzzle have been collected, but not yet put together in an integrative fashion. We think that this Research Topic facilitates structuring research and theory-building and advances models and theoretical frameworks such as the ViolEx model.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This editorial was developed in the context of the Research Training Group "Expectation Maintenance vs. Change in the Context of Expectation Violations: Connecting Different Approaches" funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Ref. no.: GRK 2271).

REFERENCES

Brandtstädter, J., and Greve, W. (1994). The aging self: stabilizing and protective processes. *Dev. Rev.* 14, 52–80. doi: 10.1006/drev.1994.1003

Gollwitzer, M., Rothmund, T., and Süssenbach, P. (2013). The Sensitivity to Mean Intentions (SeMI) model: Basic assumptions, recent findings, and potential avenues for future research. *Soc. Personal. Psychol. Compass* 7, 415–426. doi: 10.1111/spc3.12041

Greve, W., and Wentura, D. (2010). True lies: Self-stabilization without self-deception. *Conscious. Cogn.* 19, 721–730. doi: 10.1016/j.concog.2010.05.016

Hechler, T., Endres, D., and Thorwart, A. (2016). Why harmless sensations might hurt in individuals with chronic pain: about heightened prediction and perception of pain in the mind. *Front. Psychol.* 7:1638. doi: 10.3389/fpsyg.2016.01638

Müller, E. M., Makeig, S., Stemmler, G., Hennig, J., and Wacker, J. (2011). Dopamine effects on human error processing depend on COMT VAL158MET genotype. *J. Neurosci.* 31, 15818–15825. doi: 10.1523/JNEUROSCI.2103-11.2011

Müller, E. M., Panitz, C., Hermann, C., and Pizzagalli, D. A. (2014). Prefrontal oscillations during recall of conditioned and extinguished fear in humans. *J. Neurosci.* 34, 7059–7066. doi: 10.1523/JNEUROSCI.3427-13.2014

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., and Thorwart, A. (2015). Expectancies as core features of mental disorders. *Curr. Opin. Psychiatry*. 28, 378–385. doi: 10.1097/YCO.0000000000000184

Rothermund, K., and Brandtstädter, J. (2003). Coping with deficits and losses in later life: From compensatory action to accommodation. *Psychol. Aging* 18, 896–905. doi: 10.1037/0882-7974.18.4.896

Schmitt, M., Gollwitzer, M., Maes, J., and Arbach, D. (2005). Justice sensitivity: assessment and location in the personality space. *Eur. J. Psychol. Assess.* 21, 202–211. doi: 10.1027/1015-5759.21.3.202

Stinson, D. A., Logel, C., Shepherd, S., and Zanna, M. (2011). Rewriting the self-fulfilling prophecy of social rejection: self-affirmation improves relational security and social behavior up to 2 months later. *Psychol. Sci.* 22, 1145–1149. doi: 10.1177/0956797611417725

Thorwart, A., Livesey, E., Wilhelm, F., Liu, W., and Lachnit, H. (2017). Learned Predictiveness and Outcome Predictability effects are not simply two sides of the same coin. *J. Exp. Psychol. Anim. Learn. Cogn.* 43, 341–365. doi: 10.1037/xan0000150

Vlaeyen, J. W., and Linton, S. J. (2012). Fear-avoidance model of chronic musculoskeletal pain: 12 years on. *Pain* 153, 1144–1147. doi: 10.1016/j.pain.2011.12.009

Yzerbyt, V., and Carnaghi, A. (2007). "Stereotype change in the social context," in *Stereotype Dynamics: Language-Based Approaches to the Formation, Maintenance, and Transformation of Stereotypes,* eds Y. Kashima, K. Fiedler, and P. Freytag (New York: Erlbaum), 29–58.

Conflict of Interest Statement: 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.

Copyright © 2018 Gollwitzer, Thorwart and Meissner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

PERSPECTIVE
published: 18 January 2017
doi: 10.3389/fpsyg.2017.00009

On the Maintenance of Expectations in Major Depression – Investigating a Neglected Phenomenon

Tobias Kube, Winfried Rief and Julia A. Glombiewski*

Department of Clinical Psychology and Psychotherapy, Philipps-University of Marburg, Marburg, Germany

In this perspective paper, we suggest that among patients suffering from major depressive disorder (MDD), dysfunctional expectations are maintained despite experiences that are contrary to these expectations. Surprisingly, this persistence of expectations in MDD has not yet been addressed by empirical studies. We argue that it is worthwhile to investigate this phenomenon with the aim of improving the treatment of MDD, and we provide a theoretical framework for understanding it. It is hypothesized that the persistence of expectations is primarily due to a process called immunization. That is, people experiencing depressive symptoms may cognitively reappraise the contradictory experience such that expectations do not need to be changed. There may be two mechanisms underlying this immunization: (1) the experience in the expectation-violating situation is considered to be an exception; or (2) the credibility of the information gained from the experience is called into question. Moreover, the maintenance of expectations may be particularly persistent if a person's expectations reflect his or her self-concept, as self-concept has been shown to be associated with future expectations. To empirically examine the hypothesized maintenance of expectations in MDD, we propose an experimental approach which could provide important implications for the treatment of MDD within cognitive behavioral therapy. We suggest that psychological interventions such as behavioral experiments should more rigorously focus on patients' appraisal of expectation-violating experiences in order to prevent immunization processes. Therapists should continuously examine whether patients' expectations were modified and should address the reasons for the maintenance of expectations.

Keywords: major depression, expectation violation, expectancy, immunization, self-concept, expectation persistence, cognitive-behavioral therapy, behavioral experiment

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Luana Colloca, University of Maryland, USA
Jens Gaab, University of Basel, Switzerland

****Correspondence:***
Tobias Kube
tobias.kube@uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *23 September 2016*
Accepted: *03 January 2017*
Published: *18 January 2017*

Citation:
Kube T, Rief W and Glombiewski JA (2017) On the Maintenance of Expectations in Major Depression – Investigating a Neglected Phenomenon. Front. Psychol. 8:9. doi: 10.3389/fpsyg.2017.00009

THE RELEVANCE OF EXPECTATIONS IN MAJOR DEPRESSION

In a clinical psychology framework, expectations[1] have been defined as future-directed cognitions that focus on the incidence or non-incidence of a specific event or experience (Kube et al., 2016). Based on the Rescorla–Wagner model (Rescorla, 1967), expectations are developed through learning processes (Cleeremans and McClelland, 1991; Colloca and Benedetti, 2009;

[1]The terms 'expectation' and 'expectancy' are often used in an interchangeable way. However, 'expectation' is more frequently used as a specific, verbalized construct whereas 'expectancies' may be present without full awareness (i.e., implicit expectancies). In this manuscript, we only use the term 'expectation.'

Colloca and Miller, 2011). Expectations have been identified to contribute substantially to clinical outcome in various medical conditions (Auer et al., 2016; Nestoriuc et al., 2016). Moreover, expectations have been shown to be one of the major components contributing to placebo and nocebo responses in clinical trials (Rief et al., 2008, 2011; Schwarz et al., 2016), and expectations can substantially enhance the effects of drug-specific components (see Kube and Rief, 2016 for a review). With regard to antidepressant clinical trials, large placebo effects have been reported (Kirsch and Sapirstein, 1998; Kirsch et al., 2002, 2008, Rief et al., 2009), and they are assumed to be mainly based on expectation mechanisms (Shedden Mora et al., 2011; Rutherford et al., 2016). Given the great impact of expectancies in clinical research, Rief et al. (2015) have discussed expectancies as core features of mental disorders (Rief et al., 2015). For major depressive disorder (MDD), there is evidence that people suffering from MDD hold situation-specific dysfunctional expectations which may be elicited by depressive core beliefs (Kube et al., 2016). Clinical observations suggest that these expectations are maintained despite experiences that are contrary to patients' expectations ("expectation violation") (Rief and Glombiewski, 2016). Surprisingly, this observed persistence of expectations in MDD has not yet been investigated in empirical studies. In this perspective article, we argue that it is worthwhile to investigate the maintenance of expectations in MDD, and we provide a theoretical framework for it with the aim of inspiring empirical research into this neglected phenomenon. This could help to develop psychological interventions aiming at enhancing expectation change and could thus substantially improve current cognitive behavioral treatment (CBT) of MDD.

Exposure therapy for the treatment of anxiety disorders has recently focused on disconfirming disorder-specific expectations by maximizing the discrepancy between patients' expectations and actual situational outcomes in expectation-violating situations, which is discussed as promising approach to modify patients' expectations and thereby reduce anxiety symptoms (Craske et al., 2014; Craske, 2015). In MDD, however, disorder-specific expectations are less obvious: people suffering from MDD often report somatic symptoms (such as sleep disturbance, loss of appetite etc.) and negative mood, but may be less aware of cognitions such as expectations (Beck, 2011). Prior research has indicated that (treatment) outcome expectations (Greenberg et al., 2006; Price et al., 2008), self-efficacy expectancies (Ludman et al., 2003; Gopinath et al., 2007; Gordon et al., 2011), and global expectations about future events (Strunk et al., 2006; Vilhauer et al., 2012) predict the course of depressive symptoms. However, situation-specific expectations resulting from depressive core beliefs have received limited attention in psychotherapy research. Similarly, CBT of MDD has primarily focused on present-focused cognitions and automatic thoughts by using cognitive and behavioral interventions (such as cognitive restructuring and behavioral experiments), while rigorously disconfirming future-directed expectations has so far received less attention. A more focused examination of patients' expectations may be advantageous for optimizing psychological interventions (Rief and Glombiewski, 2016).

This is especially important because MDD has been shown to have a high relapse rate (Judd et al., 1998; Lin et al., 1998; Solomon et al., 2000; Pintor et al., 2003; Eaton et al., 2008; Moffitt et al., 2010). According to Risch et al. (2012), relapse may be due to the reactivation of dysfunctional thoughts when confronted with new stressful events. Moreover, a substantial group of patients does not respond to usual CBT (Hofmann et al., 2012; Button et al., 2015; Beard et al., 2016). We hypothesize that the long-term efficacy of CBT could be increased by more rigorously addressing the mechanisms underlying the persistence of dysfunctional expectations. Before discussing these clinical implications, we first address in more detail the phenomenon of expectation persistence.

FRAMEWORKS FOR THE MAINTENANCE OF EXPECTATIONS IN EXPECTATION-VIOLATING SITUATIONS

Rief et al. (2015) proposed a theoretical model to explain the development and maintenance of expectations. According to this model, expectations are shaped by learning processes, as well as by social influences and individual differences. After being confronted with experiences that are contrary to one's expectations, expectations can either be changed or maintained (Rief et al., 2015). We suggest that healthy individuals are able to change their expectations after expectation-violating experiences. For instance, though many people may initially expect to fail when attempting a novel difficult task, healthy individuals may modify their expectations about future performance after receiving feedback indicating that they performed well. However, we suggest that among individuals suffering from MDD expectations are often maintained despite experiences that are contrary to their expectations. We argue that this persistence of expectations despite contradictory experiences is a core feature of MDD, and that the maintenance of expectations in MDD is due to maladaptive information processing involving a process called "immunization."

Immunization as Important Mechanism for the Persistence of Expectations

The term "immunization" was originally introduced by Brandstädter and Greve (1994) in a developmental psychology framework and needs to be distinguished from its use in a medical context. According to Brandstädter and Greve (1994), immunization serves as self-protective mechanism by reappraising experiences of loss in a self-worth stabilizing manner. In clinical psychology, however, immunization has not yet been empirically investigated, and little is known about this phenomenon. According to Rief et al. (2015), in a clinical psychology framework, immunization means that an expectation-violating experience is cognitively reappraised so that one's prior expectation is confirmed by a *post hoc* evaluation, while the contradictory experience is discounted. We suggest that there are two possible mechanisms underlying this immunization process. First, the experience gained in the expectation-violating situation may be considered to be an exception rather than the rule. For instance, a person might maintain expectations

of failure after successful experiences by thinking, "Well, I managed that, but it was an easy task." and thus reappraising the contradictory experience. Second, a person may question the credibility of the information gained in an expectation-violating situation. For instance, the expectation "Nobody will be there for me when I ask for help" may be maintained despite another person's offer of help by a reappraisal such as, "He only helped me because he wanted to get rid of me afterward. In fact, he does not like me and is not interested in how I am feeling." Both mechanisms may lead to a persistence or possibly even reinforcement of expectations via cognitive reappraisal of the contradictory experience in a way that confirms prior expectations. In addition to this immunization process, other forms of maladaptive information processing in MDD, such as cognitive distortion, selective attention or selective memory (Beck, 1963; Hammen and Krantz, 1976; Hammen, 1978; Beck et al., 1979; Krantz and Hammen, 1979; Haaga and Beck, 1995; Beck and Haigh, 2014), may contribute to the maintenance of expectations.

A Social Psychology Perspective

The idea that individuals reappraise contrary information to experience cognitive consistency is supported by research from social and cognitive psychology (Lord et al., 1979; Ross and Lepper, 1980; Frey and Rosch, 1984; Oaksford and Chater, 2007). Cognitive consistency theories and especially the theory of cognitive dissonance (Festinger, 1962) have impacted research on how individuals change cognitions and attitudes. According to Festinger (1962), cognitive dissonance is an aversive state that is generated when a person has two or more contrary cognitions. As a result, people aim to reduce this dissonance by changing one or more of the inconsistent cognitions.

Moreover, research from social and personality psychology has provided extensive evidence that a person's self-concept remains quite stable over time, as individuals selectively search for information that confirms the self-concept while denying self-concept incongruent information (Markus, 1977; Swann and Read, 1981a,b; Swann and Hill, 1982; Markus and Wurf, 1987). Hence, people seem to be prone to a "confirmation bias," and they are supposed to use "positive test strategies," meaning that one prefers to use strategies that are considered to confirm the prior hypothesis (Klayman and Ha, 1987). More specifically, McFarlin and Blascovich (1981) demonstrated in an experimental study that an individual's level of self-esteem predicts expectations about future performance, irrespective of feedback on performance. Given that MDD is associated with low self-esteem (Lewinsohn et al., 1988; Roberts and Monroe, 1992, 1994; Joiner et al., 1999; Orth et al., 2008), we suggest that self-esteem or other aspects of an individual's self-concept may be moderator variables within the immunization process. That is, the maintenance of expectations via immunization is more likely if the expectations involved are closely related to one's self-concept. For instance, the expectation "When I have to get an important task done, I will fail at it" may be particularly persistent if an individual's self-concept includes the assumption "I am not able to adequately cope with performance-related situations." This may be the case in individuals suffering from MDD, since people experiencing depressive symptoms are thought to hold dysfunctional core beliefs such as, "I am not able to get anything done" (Beck et al., 1979; Beck, 2011). **Figure 1** illustrates the suggested immunization process while taking into account the self-concept relevance of expectations.

Also, we suggest that the maintenance of self-concept related expectations is facilitated by the fact that actively modifying one's expectations is perceived as more effortful than reappraising the experience, since one thereby does not need to change one's self-concept (see also Swann and Hill, 1982). For instance, if an individual were to change the expectation, "When I have to get an important task done, I will fail at it" into "When I have to get an important task done, I will manage it," it would follow that the individual is abandoning an excuse for not exposing oneself to performance-related situations. Our clinical experiences, however, suggest that people experiencing depressive symptoms tend to use their pessimistic expectations as justification for withdrawal and avoidance (e.g., "I do not need to try that because I will fail at it anyway"). For instance, modifying one's expectation to "I will be able to manage that" may imply that one has the responsibility to overcome existing challenges and is no longer able to use expectations about failure as excuse for withdrawal and avoidance. This may threaten the self-concept against the background of past behavior, hence facilitating expectation maintenance rather than expectation change.

A Neurobiological Perspective

Expectations have been suggested to shape experiences and to affect how an individual experiences its environment (Kirsch, 1999). This idea has recently been examined by cognitive neuroscience researchers. For instance, it has been shown that prior expectations bias stimulus processing in the visual cortex (Kok et al., 2013). Additionally, research from cognitive neuroscience has indicated that expectation-violating effects (e.g., by using invalid cues) can lead to a "surprise-attention link," resulting in a shift of attention, which may hinder or facilitate learning processes (Horstmann, 2015). Given the maladaptive information processing in MDD, this bias in experiencing one's environment by prior expectations could be especially pronounced in people suffering from MDD, which could further contribute to expectation maintenance.

INVESTIGATING THE PERSISTENCE OF EXPECTATIONS

To empirically examine the hypothesized phenomenon of expectation maintenance in MDD, we propose a stepwise experimental approach (see **Table 1**). First, researchers should attempt to empirically examine the clinical observation that people suffering from MDD tend to maintain their expectations despite expectation-violating experiences. For this purpose, researchers could focus on explicit expectation regarding personal achievement (e.g., "I will be successful in working on an unknown test"), and they could ask participants to complete an unknown test which is said to be very difficult. Then,

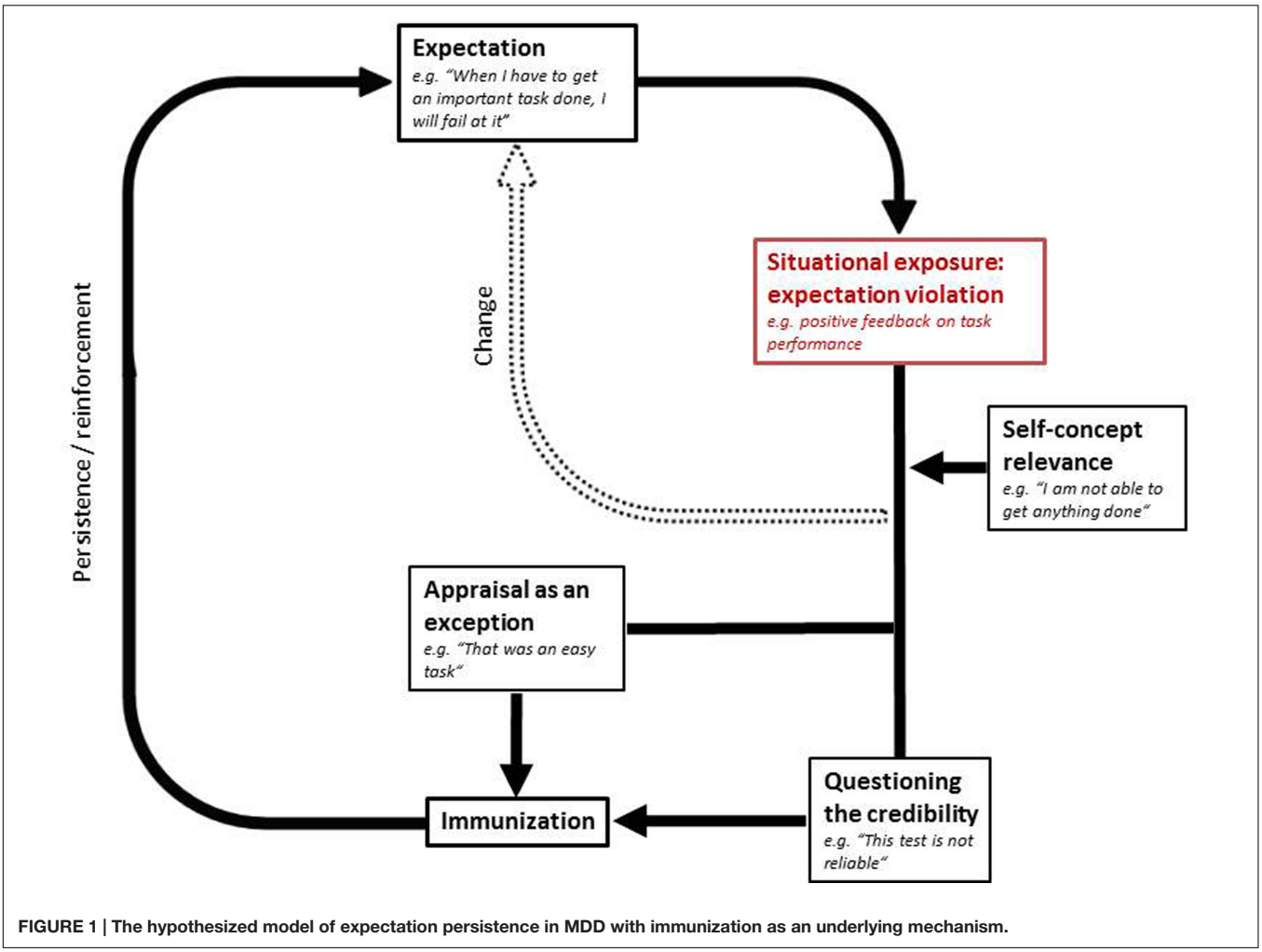


FIGURE 1 | The hypothesized model of expectation persistence in MDD with immunization as an underlying mechanism.

TABLE 1 | Proposed stepwise procedure for the investigation of expectation persistence.

	Aim of the investigation step
Step 1	Systematically observing that people suffering from MDD relative to healthy controls tend to more frequently maintain their expectations despite experiences contrary to expectations. Developing an experimental paradigm for the investigation of expectation violation in MDD. Developing a questionnaire assessing situation-specific expectations in MDD.
Step 2	Experimentally manipulating the appraisal of an expectation-violating situation and thus experimentally manipulating immunization.
Step 3	Examining the self-concept relevance of expectations as a possible moderator of immunization in correlational analyses. Subsequently, experimentally manipulating the self-concept relevance of expectations.
Step 4	Conducting a clinical study with cognitive behavior therapy enhanced with expectation focused psychological interventions vs. treatment as usual.

participants could be given standardized performance feedback that is surprisingly positive. Thereby, it could be examined whether subjects changed their initial expectations after receiving expectation-violating feedback; that is, the possible change of expectations from pre to post would be the dependent variable. At the same time, the hypothesized immunization process as an underlying mechanism could be examined by exploring the reasons for expectation change vs. expectation maintenance.

After this exploratory approach, it may be useful to experimentally manipulate the appraisal of the expectation-violating situation to impede or enhance immunization. For this purpose, experimenters could vary whether or not participants are guided to consider the expectation-violating experience as exceptional. For instance, one could provide standardized information to participants suggesting that the test completed either is or is not useful for predicting achievement in other situations. Thus, it can be examined to what degree the manipulation of the perceived relevance of the expectation-violating experience influences expectation change. Another approach for experimentally manipulating immunization could be the induction of self-focused rumination vs. distraction after an expectation-violating situation. Based on Lyubomirsky et al.'s (2003) paradigm, it is hypothesized that self-focused rumination in individuals with MDD triggers negative thoughts about perceived past failures, which may facilitate

immunization and may therefore additionally contribute to expectation maintenance. To investigate self-concept relevance as a possible moderating variable, correlational analyses could examine whether expectation maintenance is more likely if the expectations are closely related to the individual's self-concept. If correlational analyses yield promising results, researchers could experimentally vary whether or not the expectations examined in the study are associated with self-concept. Finally, clinical studies might examine whether enhancing CBT with expectation focused interventions (see also Rief and Glombiewski, 2016) increases therapy success relative to treatment as usual.

CLINICAL IMPLICATIONS

A better understanding of the persistence of expectations in MDD would have several implications for CBT for MDD. Within CBT for MDD, behavioral experiments are an effective method of testing automatic thoughts in order to facilitate cognitive restructuring (Dobson and Hamilton, 2003; Beck, 2011; Dobson, 2016). Given the relevance of disorder-specific expectations in MDD, we encourage therapists to more specifically focus on patients' expectations when designing behavioral experiments, as the "if-then" structure of expectations (as opposed to other automatic thoughts) makes them susceptible to falsification (Kube et al., 2016). That is, behavioral experiments can serve as expectation-violating situations insofar as patients can gain experiences that are contrary to their expectations (Craske et al., 2014). However, clinical experiences suggest that experiences contrary to patients' expectations do not always result in successful change of expectations (Rief and Glombiewski, 2016). In such cases, it may be worthwhile to actively explore the reasons for the maintenance of expectations in order to impede immunization processes, which could improve therapy success in multiple ways.

First, if a patient considers the experience in a behavioral experiment to be an exception, the therapist should discuss whether this appraisal is accurate or useful. If necessary, behavioral experiments may subsequently be repeated under different circumstances to call the patient's appraisal into question. Thus, the generalizability of the experience gained in a behavioral experiment should be emphasized to prevent immunization processes. Second, if a patient fundamentally questions the credibility of the experience, the therapist might help the patient to re-examine the validity of the experience. Third, therapists should carefully consider whether the expectations tested in a behavioral experiment are closely related to the patient's self-concept, and should be aware that if so, change in expectations may be less likely. Such awareness may prevent disappointment for both patient and therapist, and the therapist can motivate the patient to change his or her behavior, e.g., by discussing the consequences of the behavior. Fourth, in addition to exploring the reasons for maintenance of expectations *after* a behavioral experiment, it may be useful to discuss with the patient the conditions under which he/she would change his/her expectations *before* engaging in the behavioral experiment. This would allow the therapist and patient to agree on the conditions for the behavioral experiment such that the patient would consider a violation of his/her expectations to be a valid experience. This procedure might help to prevent *post hoc* confirmation of expectations via immunization.

Given the high relapse rates in MDD (Judd et al., 1998; Lin et al., 1998; Solomon et al., 2000; Pintor et al., 2003; Eaton et al., 2008; Moffitt et al., 2010), rigorously addressing patients' expectations may be helpful with respect to long-term benefit from therapy, as patients can be encouraged to test future dysfunctional expectations independently after therapy completion. If CBT were to enable patients to prevent dysfunctional immunization processes, this could result in additional positive experiences which in turn could impede the reactivation of dysfunctional thoughts (Risch et al., 2012).

Considering the maintenance of expectations may also be useful for the treatment of other mental disorders. Modifying patients' expectations through exposure to expectation-violating situations has been discussed as a promising approach in the treatment of anxiety disorders (Craske et al., 2014; Craske, 2015), obsessive compulsive disorders (Craske et al., 2014), and chronic pain (Riecke et al., 2013). We believe that impeding immunization processes (as discussed for MDD in this article) might also be an important mechanism of change in these disorders. Thus, we hope that the proposed theoretical model for the persistence of expectations will inspire future research with the aim of optimizing cognitive-behavioral therapy by preventing immunization processes not only in MDD, but also in other mental disorders involving dysfunctional expectations.

CONCLUSION

The maintenance of expectations despite experiences that are contrary to expectations is believed to be a core feature of MDD. We suggest that this persistence of expectations is due to maladaptive information processing in MDD, in particular, immunization processes. Immunization is hypothesized to be especially pronounced if an individual's expectations are closely associated with his or her self-concept. This should be examined in a series of experimental studies and could provide useful information for the treatment of depression. Carefully addressing the reasons for expectation persistence may be useful for optimizing psychological interventions, hence increasing the long-term efficacy of CBT.

AUTHOR CONTRIBUTIONS

TK: Did the major part of the work with regard to conception and design; mainly contributed to the development of the manuscript; approves the manuscript to be published; agrees on being accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

WR: Substantially contributed to the conception of the work; revised the manuscript critically for important intellectual content; approves the manuscript to be published; agrees on being accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. JG: Substantially contributed to the conception of the work; revised the manuscript critically for important intellectual content; approves the manuscript to be published; agrees on being accountable for all aspects of the work in ensuring that.

ACKNOWLEDGMENT

The authors thank Ashley Witt for her helpful correction of the manuscript.

REFERENCES

Auer, C. J., Glombiewski, J. A., Doering, B. K., Winkler, A., Laferton, J. A. C., Broadbent, E., et al. (2016). Patients' expectations predict surgery outcomes: a meta-analysis. *Int. J. Behav. Med.* 23, 49–62. doi: 10.1007/s12529-015-9500-4

Beard, C., Stein, A. T., Hearon, B. A., Lee, J., Hsu, K. J., and Bjorgvinsson, T. (2016). Predictors of depression treatment response in an intensive CBT partial hospital. *J. Clin. Psychol.* 72, 297–310. doi: 10.1002/jclp.22269

Beck, A., Rush, A., Shaw, B., and Emery, G. (1979). *Cognitive Therapy of Depression*. New York, NY: Guilford Press.

Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Arch. Gen. Psychiatry* 9, 324–333. doi: 10.1001/archpsyc.1963.01720160014002

Beck, A. T., and Haigh, E. A. P. (2014). Advances in cognitive theory and therapy: the generic cognitive model. *Annu. Rev. Clin. Psychol.* 10, 1–24. doi: 10.1146/annurev-clinpsy-032813-153734

Beck, J. S. (2011). *Cognitive Behavior Therapy: Basics and Beyond*. New York, NY: Guilford Press.

Brandstädter, J., and Greve, W. (1994). The aging self: stabilizing and protective processes. *Dev. Rev.* 14, 52–80. doi: 10.1006/drev.1994.1003

Button, K. S., Turner, N., Campbell, J., Kessler, D., Kuyken, W., Lewis, G., et al. (2015). Moderators of response to cognitive behavioural therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care. *J. Affect. Disord.* 174, 272–280. doi: 10.1016/j.jad.2014.11.057

Cleeremans, A., and McClelland, J. L. (1991). Learning the structure of event sequences. *J. Exp. Psychol. Gen.* 120, 235–253. doi: 10.1037/0096-3445.120.3.235

Colloca, L., and Benedetti, F. (2009). Placebo analgesia induced by social observational learning. *Pain* 144, 28–34. doi: 10.1016/j.pain.2009.01.033

Colloca, L., and Miller, F. G. (2011). How placebo responses are formed: a learning perspective. *Philos. Trans. R. Soc. B Biol. Sci.* 366, 1859–1869. doi: 10.1098/rstb.2010.0398

Craske, M. (2015). Optimizing exposure therapy for anxiety disorders: an inhibitory learning and inhibitory regulation approach. *Verhaltenstherapie* 25, 134–143. doi: 10.1159/000381574

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., and Vervliet, B. (2014). Maximizing exposure therapy: an inhibitory learning approach. *Behav. Res. Ther.* 58, 10–23. doi: 10.1016/j.brat.2014.04.006

Dobson, K. S. (2016). New frontiers in cognitive-behavioral therapy for depression. *Int. J. Cogn. Ther.* 9, 107–123. doi: 10.1521/ijct.2016.9.2.107

Dobson, K. S., and Hamilton, K. E. (2003). "Cognitive restructuring: behavioral tests of negative cognitions," in *Cognitive Behavior Therapy: Applying Empirically Supported Techniques in your Practice*, eds J. E. F. W. O'Donohue and S. C. Hayes (New York, NY: Wiley), 84–88.

Eaton, W. W., Shao, H., Nestadt, G., Lee, B. H., Bienvenu, O. J., and Zandi, P. (2008). Population-based study of first onset and chronicity in major depressive disorder. *Arch. Gen. Psychiatry* 65, 513–520. doi: 10.1001/archpsyc.65.5.513

Festinger, L. (1962). *A Theory of Cognitive Dissonance*. Palo Alto CA: Stanford university press.

Frey, D., and Rosch, M. (1984). Information seeking after decisions the roles of novelty of information and decision reversibility. *Personal. Soc. Psychol. Bull.* 10, 91–98. doi: 10.1177/0146167284101010

Gopinath, S., Katon, W. J., Russo, J. E., and Ludman, E. J. (2007). Clinical factors associated with relapse in primary care patients with chronic or recurrent depression. *J. Affect. Disord.* 101, 57–63. doi: 10.1016/j.jad.2006.10.023

Gordon, M. S., Tonge, B., and Melvin, G. A. (2011). Outcome of adolescent depression: 6 months after treatment. *Aust. N. Z. J. Psychiatry* 45, 232–239. doi: 10.3109/00048674.2010.538838

Greenberg, R. P., Constantino, M. J., and Bruce, N. (2006). Are patient expectations still relevant for psychotherapy process and outcome? *Clin. Psychol. Rev.* 26, 657–678. doi: 10.1016/j.cpr.2005.03.002

Haaga, D. A. F., and Beck, A. T. (1995). Perspectives on depressive realism–implications for cognitive theory of depression. *Behav. Res. Ther.* 33, 41–48. doi: 10.1016/0005-7967(94)e0016-c

Hammen, C., and Krantz, S. (1976). Effect of success and failure on depressive cognitions. *J. Abnorm. Psychol.* 85, 577–586. doi: 10.1037/0021-843X.85.6.577

Hammen, C. L. (1978). Depression, distortion, and life stress in college students. *Cogn. Ther. Res.* 2, 189–192. doi: 10.1007/BF01172733

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., and Fang, A. (2012). The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cogn. Ther. Res.* 36, 427–440. doi: 10.1007/s10608-012-9476-1

Horstmann, G. (2015). The surprise-attention link: a review. *Ann. N. Y. Acad. Sci.* 1339, 106–115. doi: 10.1111/nyas.12679

Joiner, T. E., Katz, J., and Lew, A. (1999). Harbingers of depressotypic reassurance seeking: negative life events, increased anxiety, and decreased self-esteem. *Personal. Soc. Psychol. Bull.* 25, 630–637. doi: 10.1177/0146167299025005008

Judd, L. L., Akiskal, H. S., Maser, J. D., Zeller, P. J., Endicott, J., Coryell, W., et al. (1998). A prospective 12-year study of subsyndromal and syndromal depressive symptoms in unipolar major depressive disorders. *Arch. Gen. Psychiatry* 55, 694–700. doi: 10.1001/archpsyc.55.8.694

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., and Johnson, B. T. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 5:e45. doi: 10.1371/journal.pmed.0050045

Kirsch, I., Moore, T. J., Scoboria, A., and Nicholls, S. S. (2002). The emperor's new drugs: an analysis of antidepressant medication data submitted to the US Food and Drug Administration. *Prev. Treat.* 5, 23. doi: 10.1016/j.ijlp.2008.09.001

Kirsch, I., and Sapirstein, G. (1998). Listening to Prozac but hearing placebo: a meta-analysis of antidepressant medication. *Prev. Treat.* 1, 2. doi: 10.1037/1522-3736.1.1.12a

Kirsch, I. E. (1999). *How Expectancies Shape Experience*. Washington DC: American Psychological Association.

Klayman, J., and Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* 94, 211–228. doi: 10.1037/0033-295X.94.2.211

Kok, P., Brouwer, G. J., van Gerven, M. A. J., and de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *J. Neurosci.* 33, 16275–16284. doi: 10.1523/jneurosci.0742-13.2013

Krantz, S., and Hammen, C. (1979). Assessment of cognitive bias in depression. *J. Abnorm. Psychol.* 88, 611–619. doi: 10.1037/0021-843X.88.6.611

Kube, T., D'Astolfo, L., Glombiewski, J. A., Doering, B. K., and Rief, W. (2016). Focusing on situation-specific expectations in major depression as basis for behavioral experiments - development of the depressive expectations scale. *Psychol. Psychother.* doi: 10.1111/papt.12114 [Epub ahead of print].

Kube, T., and Rief, W. (2016). Are placebo and drug-specific effects additive? – Questioning basic assumptions of double-blinded randomized clinical trials and presenting novel study designs. *Drug Discov. Today* doi: 10.1016/j.drudis.2016.11.022 [Epub ahead of print].

Lewinsohn, P. M., Hoberman, H. M., and Rosenbaum, M. (1988). A prospective study of risk factors for unipolar depression. *J. Abnorm. Psychol.* 97, 251–264. doi: 10.1037/0021-843x.97.3.251

Lin, E. H. B., Katon, W. J., VonKorff, M., Russo, J. E., Simon, G. E., Bush, T. M., et al. (1998). Relapse of depression in primary care–Rate and clinical predictors. *Arch. Fam. Med.* 7, 443–449. doi: 10.1001/archfami.7.5.443

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* 37, 2098–2109. doi: 10.1037/0022-3514.37.11.2098

Ludman, E., Katon, W., Bush, T., Rutter, C., Lin, E., Simon, G., et al. (2003). Behavioural factors associated with symptom outcomes in a primary care-based depression prevention intervention trial. *Psychol. Med.* 33, 1061–1070. doi: 10.1017/s003329170300816x

Lyubomirsky, S., Kasri, F., and Zehm, K. (2003). Dysphoric rumination impairs concentration on academic tasks. *Cogn. Ther. Res.* 27, 309–330. doi: 10.1023/a:1023918517378

Markus, H. (1977). Self-schemata and processing information about the self. *J. Pers. Soc. Psychol.* 35, 63–78. doi: 10.1037//0022-3514.35.2.63

Markus, H., and Wurf, E. (1987). The dynamic self-concept: a social psychological perspective. *Annu. Rev. Psychol.* 38, 299–337. doi: 10.1146/annurev.psych.38.1.299

McFarlin, D. B., and Blascovich, J. (1981). Effects of self-esteem and performance feedback on future affective preferences and cognitive expectations. *J. Pers. Soc. Psychol.* 40, 521–531. doi: 10.1037/0022-3514.40.3.521

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., et al. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychol. Med.* 40, 899–909. doi: 10.1017/s0033291709991036

Nestoriuc, Y., von Blanckenburg, P., Schuricht, F., Barsky, A. J., Hadji, P., Albert, U. S., et al. (2016). Is it best to expect the worst? Influence of patients' side-effect expectations on endocrine treatment outcome in a 2-year prospective clinical cohort study. *Ann. Oncol.* 27, 1909–1915. doi: 10.1093/annonc/mdw266

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Orth, U., Robins, R. W., and Roberts, B. W. (2008). Low self-esteem prospectively predicts depression in adolescence and young adulthood. *J. Pers. Soc. Psychol.* 95, 695–708. doi: 10.1037/0022-3514.95.3.695

Pintor, L., Gasto, C., Navarro, V., Torres, X., and Fananas, L. (2003). Relapse of major depression after complete and partial remission during a 2-year follow-up. *J. Affect. Disord.* 73, 237–244. doi: 10.1016/s0165-0327(01)00480-3

Price, M., Anderson, P., Henrich, C. C., and Rothbaum, B. O. (2008). Greater expectations: using hierarchical linear modeling to examine expectancy for treatment outcome as a predictor of treatment response. *Behav. Ther.* 39, 398–405. doi: 10.1016/j.beth.2007.12.002

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychol. Rev.* 74, 71–80. doi: 10.1037/h0024109

Riecke, J., Holzapfel, S., Rief, W., and Glombiewski, J. A. (2013). Evaluation and implementation of graded in vivo exposure for chronic low back pain in a German outpatient setting: a study protocol of a randomized controlled trial. *Trials* 14, 203. doi: 10.1186/1745-6215-14-203

Rief, W., Bingel, U., Schedlowski, M., and Enck, P. (2011). Mechanisms involved in placebo and nocebo responses and implications for drug trials. *Clin. Pharmacol. Ther.* 90, 722–726. doi: 10.1038/clpt.2011.204

Rief, W., and Glombiewski, J. A. (2016). Expectation-Focused Psychological Interventions (EFPI). *Verhaltenstherapie* 26, 47–54. doi: 10.1159/000442374

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubo, A., Schwarting, R., and Thorwart, A. (2015). Expectancies as core features of mental disorders. *Curr. Opin. Psychiatry* 28, 378–385. doi: 10.1097/yco.0000000000000184

Rief, W., Hofmann, S. G., and Nestoriuc, Y. (2008). The power of expectation–Understanding the placebo and nocebo phenomenon. *Soc. Personal. Psychol. Compass* 2, 1624–1637. doi: 10.1111/j.1751-9004.2008.00121.x

Rief, W., Nestoriuc, Y., Weiss, S., Welzel, E., Barsky, A. J., and Hofmann, S. G. (2009). Meta-analysis of the placebo response in antidepressant trials. *J. Affect. Disord.* 118, 1–8. doi: 10.1016/j.jad.2009.01.029

Risch, A. K., Stangier, U., Heidenreich, T., and Hautzinger, M. (2012). *Kognitive Erhaltungstherapie bei Rezidivierender Depression*. Berlin: Springer.

Roberts, J. E., and Monroe, S. M. (1992). Vulnerable self-esteem and depressive symptoms: prospective findings comparing three alternative conceptualizations. *J. Pers. Soc. Psychol.* 62, 804–812. doi: 10.1037//0022-3514.62.5.804

Roberts, J. E., and Monroe, S. M. (1994). A multidimensional model of self-esteem in depression. *Clin. Psychol. Rev.* 14, 161–181. doi: 10.1016/0272-7358(94)90006-x

Ross, L., and Lepper, M. R. (1980). "The perseverance of beliefs: empirical and normative considerations," in *New Directions for Methodology of Behavioral Science: Fallible Judgment in Behavioral Research*, ed. R. A. Shweder (San Francisco, CA: Jossey-Bass), 17–36.

Rutherford, B. R., Wall, M. M., Brown, P. J., Choo, T.-H., Wager, T. D., Peterson, B. S., et al. (2016). Patient expectancy as a mediator of placebo effects in antidepressant clinical trials. *Am. J. Psychiatry* doi: 10.1176/appi.ajp.2016.16020225 [Epub ahead of print].

Schwarz, K. A., Pfister, R., and Buchel, C. (2016). Rethinking explicit expectations: connecting placebos, social cognition, and contextual perception. *Trends Cogn. Sci.* 20, 469–480. doi: 10.1016/j.tics.2016.04.001

Shedden Mora, M., Nestoriuc, Y., and Rief, W. (2011). Lessons learned from placebo groups in antidepressant trials. *Philos. Trans. R. Soc. B Biol. Sci.* 366, 1879–1888. doi: 10.1098/rstb.2010.0394

Solomon, D. A., Keller, M. B., Leon, A. C., Mueller, T. I., Lavori, P. W., Shea, T., et al. (2000). Multiple recurrences of major depressive disorder. *Am. J. Psychiatry* 157, 229–233. doi: 10.1176/appi.ajp.157.2.229

Strunk, D. R., Lopez, H., and DeRubeis, R. J. (2006). Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behav. Res. Ther.* 44, 861–882. doi: 10.1016/j.brat.2005.07.001

Swann, W. B., and Hill, C. A. (1982). When our identities are mistaken: reaffirming self-conceptions through social interaction. *J. Pers. Soc. Psychol.* 43, 59–66. doi: 10.1037//0022-3514.43.1.59

Swann, W. B., and Read, S. J. (1981a). Acquiring self-knowledge: the search for feedback that fits. *J. Pers. Soc. Psychol.* 41, 1119–1128. doi: 10.1037/0022-3514.41.6.1119

Swann, W. B., and Read, S. J. (1981b). Self-verification processes: how we sustain our self-conceptions. *J. Exp. Soc. Psychol.* 17, 351–372. doi: 10.1016/0022-1031(81)90043-3

Vilhauer, J. S., Young, S., Kealoha, C., Borrmann, J., IsHak, W. W., Rapaport, M. H., et al. (2012). Treating major depression by creating positive expectations for the future: a pilot study for the effectiveness of Future-Directed Therapy (FDT) on symptom severity and quality of life. *CNS Neurosci. Ther.* 18, 102–109. doi: 10.1111/j.1755-5949.2011.00235.x

Conflict of Interest Statement: 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.

Copyright © 2017 Kube, Rief and Glombiewski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

MINI REVIEW
published: 28 November 2016
doi: 10.3389/fpsyg.2016.01876

Can Psychological Expectation Models Be Adapted for Placebo Research?

Winfried Rief[1]* *and Keith J. Petrie*[2]

[1] Department of Psychology, Philipps University Marburg, Marburg, Germany, [2] Department of Psychological Medicine, University of Auckland, Auckland, New Zealand

Placebo responses contribute substantially to the effect and clinical outcome of medical treatments. Patients' expectations have been identified as one of the major mechanisms contributing to placebo effects. However, to date a general theoretical framework to better understand how patient expectations interact with features of medical treatment has not been developed. In this paper we outline an expectation model that can be used as framework for experimental studies on both placebo and nocebo mechanisms. This model is based on psychological concepts of expectation development, expectation maintenance, and expectation change within the typical paradigms used in placebo research. This theoretical framework reflects the dynamic aspects of the interaction between expectations and medical treatment, and offers a platform to combine psychological and neurophysiological research activities. Moreover, this model can be used to identify important future research questions. For example, we argue that the dynamic processes of expectation maintenance vs. expectation changes are not sufficiently addressed in current research on placebo mechanisms. Therefore, the question about how to change and optimize patients' expectations prior to treatment should be a special focus of future clinical research.

Keywords: expectation, placebo, nocebo, prediction error, expectation violation, associative learning

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Roland Thomaschke, University of Regensburg, Germany
Robert Gaschler, FernUniversität in Hagen, Germany

***Correspondence:**
Winfried Rief
rief@uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *20 September 2016*
Accepted: *14 November 2016*
Published: *28 November 2016*

Citation:
Rief W and Petrie KJ (2016) Can Psychological Expectation Models Be Adapted for Placebo Research? Front. Psychol. 7:1876. doi: 10.3389/fpsyg.2016.01876

INTRODUCTION

Placebo mechanisms contribute substantially to clinical outcome in many fields of medicine (Schedlowski et al., 2015). In randomized clinical trials, patients receiving placebo treatment typically achieve results that are almost equivalent to the response of the active intervention group. This has been shown not only for patient reported outcomes, such as pain and depression, but also for objectively assessed biological parameters such as immune reactions (Schedlowski et al., 2015), cardiovascular reactions (Meissner, 2008), or polysomnographic assessments of pain and sleep variables (Winkler and Rief, 2015).

Expectations have been identified as one of the major components contributing to placebo reactions (Schwarz et al., 2016). If patients have a need for medical interventions, they are exposed to stimuli in the clinical setting that trigger specific treatment- and outcome expectations. These stimuli include the nature of the treatment itself—such as surgery, medicines, or injections. They also include the characteristics of the clinician and the relationship formed with the patient as well as the doctor's confidence in the therapy and explanation of the treatment. The wider treatment context such as the reputation of the facility and status of the clinic may also impact

on treatment outcome expectations. As these are all factors that operate psychologically to enhance or decrease the placebo response, expectation theories can contribute to a better understanding of placebo effects. In this paper, we will use the terms expectation and expectancy interchangeable, although expectancy is more frequently used when also including implicit expectations and implicit expectation effects.

Atkinson's expectancy-value theory outlines that behavior in challenging situations is predicted by the interaction of prior expectations to be able to manage such a challenge successfully and the subjective value of the specific task (Atkinson and Reitman, 1956). In the health setting, the value of the challenge is typically associated with the hope to survive the illness and to reduce the burden caused by its pain and symptoms. According to the theory, a better clinical outcome is predicted if the expected improvements caused by a treatment are of high personal value and patients have a strong self-belief to be able to cope with the situation (self-efficacy). Indeed, low expectations of specific self-efficacy, and low expectations of therapy-driven improvements result in low treatment adherence (Horne and Weinman, 2002).

A further relevant background theory is "prospect theory" developed by Kahneman and Tversky (1979). This theory emphasizes the subjectivity of the definition by which an outcome can be considered as gains vs. losses. The authors highlight the fact that potential losses are frequently more relevant for behavioral decisions than expected gains. Applying this theory to the clinical context, patients' anxieties, and concerns about treatment can be more relevant to predict their behavior than the expected benefits of their treatment.

Expectations are frequently developed through a process of associative learning. An important model predicting how repeated trials of associative learning can lead to learned reactions is Rescorla-Wagner's model, which has been principally developed to explain Pavlovian conditioning effects (Rescorla, 1967). This model has also substantial relevance for understanding the development and the consequences of expectations. The power of an expectation corresponds in part to the associative strength in the formula of this model. Accordingly, the strength of expectation is dependent on the number of trials confirming these associations and/or the learning rate. Additionally, the model also postulates that expectations can eventually achieve a maximum level that limits further increases in association. Learning is reconceptualized as a change of expectations. Therefore, the discrepancy between expected outcome and experienced outcome is a major precondition to initiate learning processes. The important contribution of this model is for the understanding how expectations are modified. The Rescorla-Wagner Model became one of the basic concepts that stimulated the development of paradigms investigating prediction and prediction error effects in neuroscience (Schwarz et al., 2016).

While this selective collection of psychological theories on expectation is not comprehensive, it illustrates that these psychological theories have been developed with a strong non-clinical focus. Therefore, we want to develop a theoretical framework for expectation effects in the clinical context, that offers a platform to integrate these psychological theories with empirical approaches that will help explain placebo and nocebo effects in the context of medical treatments.

ADAPTING THE VIOLEX-MODEL

Recently, we developed a general model that conceptualizes how expectations influence various outcomes in clinical psychology, and when expectation violations lead to a change vs. a persistence of expectations ("the ViolEx-model;" Rief et al., 2015). The original model was developed as a broad theoretical framework to better understand the dynamic interactions between expectation effects, expectation violations, and their feedback loops to modify expectations in general. Here, we adapt this model to placebo and nocebo research and clinical encounters.

The core of the model in **Figure 1** is the interaction of expectations and clinical situations, such as visiting a doctor for the treatment of bothersome symptoms. This interaction results in predictions, outcome, and outcome evaluations that either confirm or disconfirm pre-existing expectations. The model is complemented by adding trait factors, past learning processes, and state factors to better understand how expectations developed. Different aspects of the model are covered below.

Placebo effects occur when a medical treatment and its context trigger specific expectations about a positive therapeutic outcome. Pre-existing optimistic expectations can amplify the positive effects of treatments (placebo effects), but negative expectations can also induce adverse treatment effects, such as side effects or the absence of treatment-typical improvements (nocebo effects).

The interaction of pre-existing generalized expectations and medical setting variables leads to situations-specific predictions that are associated with typical anticipatory reactions. When a treatment outcome is perceived, an individual evaluates whether it corresponds to the predicted outcome, or whether the outcome is unpredicted, such as when side effects occur. The more frequently expected positive outcomes occur then the generalized expectations are more stable, although this learning process is asymptomatic according to the Rescorla-Wagner-Model ("confirmation," see **Figure 1**). If the expected outcome does not occur, or additional unexpected outcomes develop, this will typically lead to a modification of expectations due to expectation-violating experiences ("modification," see **Figure 1**). However, it would not be adaptive if individuals were to change their expectations just because of one disconfirming event.

In reality, many people stick to their expectations despite contradictory experiences (e.g., persistence of stereotypes about population groups despite positive experiences with members of them). In the clinical context, the change of expectations is a crucial aspect, although this aspect has been poorly investigated. Patients do not show up in clinical settings without any treatment expectations, but these are not fully concordant with what doctors would like them to expect about their treatment. Therefore, it is quite typical that there is a conflict between patients' expectations (and fears) of a treatment vs. doctors' beliefs about the same therapy. Effects of self-generated

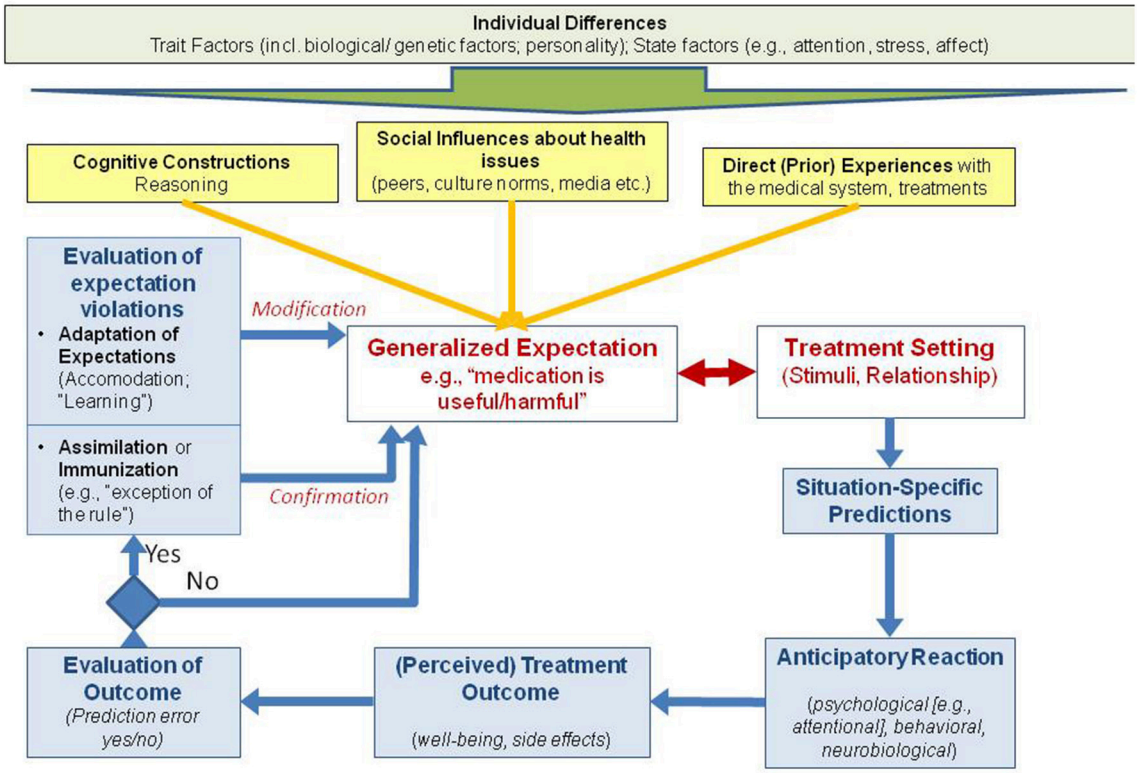


FIGURE 1 | Adaptation of the VIOL-EX model for placebo research.

expectations are usually stronger than expectations induced from outside the individual (Acosta, 1982; Kemper et al., 2012; Gaschler et al., 2014). The clinical task is thus not to establish new expectations in "naïve" patients, but to change and optimize pre-existing treatment expectations in patients.

Three factors contribute in particular to the development of expectations (see **Figure 1**, yellow connections). These are prior experience with the health care system (associative learning), social influences about health issues that are established via prior observations or learned indirectly from significant others or through media sources such as the internet. The third process that contributes is the individual personal construction of assumptions as well as the direct instructions received from others. As an example, observational learning is also of central relevance in the clinical context. Patients often have contact or observe other patients, be it in the waiting room of an outpatient clinic, or in a typical inpatient setting. These other patients can either praise or model the improvements from treatment, discuss the skill of a particular doctor, or they can complain about unwanted effects of interventions. The observation of such behavior has been shown to influence the results of the observing patient's treatment (Colloca and Benedetti, 2009; Voegtle et al., 2013; Faasse et al., 2015a).

Most of the associations indicated in **Figure 1** are also influenced by pre-existing trait factors (e.g., genetic factors, personality factors), but also by state factors such as selective attention or current options for memory retrieval. Expectancy discrepant effects can lead to a "surprise-attention link" with a shift of attention, which can facilitate or hinder learning processes (Horstmann, 2015).

The "individual differences" mentioned on top of **Figure 1** should be interpreted as a dimension influencing most other processes on all levels of this model. The effect of expectations can be also different depending whether they are self-generated vs. cue-induced expectations (Gaschler et al., 2014), with physician's interventions representing more cue-induced expectations. In part, this can help to explain why some physician-induced expectations are less powerful than patient-generated expectations.

EXAMPLES OF EMPIRICAL RESULTS TO COMPONENTS OF THE EXPECTATION MODEL

The most simple way to induce specific patients' expectations is by offering instructions about expected outcomes. In placebo research, this is typically done by informing patients that a placebo intervention is supposed to be a pain killer (Pollo et al., 2001; Bingel et al., 2011). This effect can be further amplified by inducing positive prior experiences with this specific treatment. Manipulated feedback can also induce expectations that (placebo) treatments can induce strong intervention effects.

Associative learning paradigms using Pavlovian conditioning have been used to demonstrate influences on expectations, not only in pain (Colloca and Benedetti, 2006), but also in various other conditions. Using a similar design, we were able to show that patients can "learn" to develop side effects if they received several applications with the antidepressant amitriptyline, even if eventually the drug is switched to a placebo pill (Rheker et al., 2016). Further, many people have learned that effective drugs are associated with some side effects that indicate the drug is working or powerful. This led to work showing that so called "active placebos" simulating drug-typical side effects induce more powerful placebo responses than "passive placebos" (Moncrieff et al., 2004; Rief and Glombiewski, 2012; Benedetti et al., 2013).

Generalized expectations about medical treatments are not only able to predict positive outcome, but also to predict the development of side effects and other negative outcomes (Faasse and Petrie, 2013). Promoting negative expectations can even abolish the pain-relieving effects of powerful opioids, such as remifentanil (Bingel et al., 2011). Negative beliefs about medicine predict the development of more side effects (Nestoriuc et al., 2010). This can take the form of a general belief that an individual is highly sensitive to the effects of medication in general or sensitive to specific type of medication (Horne et al., 2013; Faasse et al., 2015b). For example, expectations about developing medication side effects for endocrine therapy following breast cancer predicts more problems after treatment onset (Nestoriuc et al., 2016).

The context and environment that the medical treatment is administered is of relevance, in particular, if it includes stimuli that can activate expectations because of prior stimuli-associated experiences. The treatment context can further amplify the effect of positive expectations, e.g., if it is considered to be very professional, friendly, and clean. Treatment-context conditions are also able to influence the reactions to antidepressant drugs, and can even trigger negative effects of antidepressants compared to placebo (Rief et al., 2016). A special aspect of the treatment context is the relationship between therapist and patient. While a positive therapeutic relationship can predict successful treatment outcome, a negative therapeutic relationships can also facilitate the development of adverse treatment effects (Kaptchuk et al., 2008; Koudriavtseva et al., 2012). Moreover, the quality of the therapeutic relationship further predicts patients' adherence, and this association can also contribute to a positive outcome.

In experimental pain research, it has been shown that situation-specific predictions of pain or pain relief activate brain areas that facilitate the expected perceptions (Koyama et al., 2005). When selecting actions (such as drug intake), the brain pre-activates the representation of the predicted consequences (Waszak et al., 2012).

Further biological and psychological pathways of action of specific intervention predictions have been described (Schedlowski et al., 2015). Of particular relevance is also the role of selective attention. If patients expect adverse experiences, they also focus their attention to the specific side effect expected, which increases the perception intensity and facilitates the reporting of adverse outcomes in general (Barsky et al., 2002) or specific to the type of expectations generated (Crichton et al., 2014). Attentional processes themselves can be the result of learning (Mackintosh, 1975; Kruschke, 2003).

If the outcome is as positive as expected, this leads to a confirmation of expectations consistent with the Rescorla-Wagner Model; there is no change in association strength, hence no learning. However, if expectations are not confirmed, it remains unclear how the person will deal with that fact. Several treatment approaches actually set out to induce expectation violations (e.g., exposure therapy in anxiety disorders; Craske et al., 2014; Craske, 2015). However, not every expectation violation subsequently leads to expectation changes. Frequently, patients activate cognitive-attributional assimilation or immunization strategies to weaken or eliminate the expectation violation. The result of successful exposure sessions and other intended expectation violations can be devaluated with cognitions such as: "this was the exception to the rule;" "this only works if a therapist is close to me." A side-effect free day can still confirm side effect expectations via attributions like: "if I didn't get side effects today, I will probably get them tomorrow." While these assimilation and immunization processes have been extensively studied in social psychology, an examination of their role in clinical research is still in a very early stage.

The dynamic process of expectation development, maintenance, and change in the clinical context is further influenced by biological and psychological trait and state factors. Genetic aspects can predict whether a person is prone to develop side effects (Wendt et al., 2014), as well as whether a person is prone to develop placebo responses (Hall et al., 2012). Anxiety as a personality factor is able to predict the development of somatic symptoms as a reaction to medical interventions, but also has potential as a predictor of symptom development caused by expected environmental influences (Petrie et al., 2005; Page et al., 2006; Witthöft and Rubin, 2013; Crichton et al., 2014). The current level of biological stress reactions can further influence the interaction of the components of our model. These are just a few examples that the model presented in **Figure 1**, although already elaborated, is still an approximation, and simplification of the various influences that determine the interaction between expectations and treatment settings.

CONCLUSION

The effect of the interaction between patients' expectations and treatment context depends on past experiences, and they are characterized by dynamic interactions that happen during and after the treatment encounter. Most components are also influenced by biological and psychological individual differences such as genetic, personality, and state factors. In total, this model offers a theoretical framework that helps to communicate and connect the different approaches on placebo and nocebo research, both on a more basic scientific level and in terms of clinical applications. It also helps to identify research areas needing more work. One of the specific conclusions we want to draw is that more research is needed how to modify pre-existing expectations in situations

where patients' pre-treatment expectations are non-adaptive. Therefore, the focus of research has to move from mere inductions of specific expectations to a better understanding of processes of expectation persistence, expectation violation, and expectation change.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct, and intellectual contribution to the work, and approved it for publication. Both authors contributed equally to this manuscript.

REFERENCES

Acosta, E. Jr. (1982). Subjective and objective determinants of expectancy: similarities and differences. *Am. J. Psychol.* 95, 139–160.

Atkinson, J. W., and Reitman, W. R. (1956). Performance as a function of motive strength and expectancy of goal-attainment. *J. Abnorm. Soc. Psychol.* 53, 361–366. doi: 10.1037/h0043477

Barsky, A. J., Saintfort, R., Rogers, M. P., and Borus, J. F. (2002). Nonspecific medication side effects and the nocebo phenomenon. *JAMA* 287, 622–626. doi: 10.1001/jama.287.5.622

Benedetti, F., Thoen, W., Blanchard, C., Vighetti, S., and Arduino, C. (2013). Pain as a reward: changing the meaning of pain from negative to positive co-activates opioid and cannabinoid systems. *Pain* 154, 361–367. doi: 10.1016/j.pain.2012.11.007

Bingel, U., Wanigasekera, V., Wiech, K., Ni Mhuircheartaigh, R., Lee, M. C., and Tracey, I. (2011). The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid Remifentanil. *Sci. Transl. Med.* 3, 70ra14. doi: 10.1126/scitranslmed.3001244

Colloca, L., and Benedetti, F. (2006). How prior experience shapes placebo analgesia. *Pain* 124, 126–133. doi: 10.1016/j.pain.2006.04.005

Colloca, L., and Benedetti, F. (2009). Placebo analgesia induced by social observational learning. *Pain* 144, 28–34. doi: 10.1016/j.pain.2009.01.033

Craske, M. G. (2015). Optimizing exposure therapy for anxiety disorders: an inhibitory learning and inhibitory regulation approach. *Verhaltenstherapie* 25, 134–144. doi: 10.1159/000381574

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., and Vervliet, B. (2014). Maximizing exposure therapy: an inhibitory learning approach. *Behav. Res. Ther.* 58, 10–23. doi: 10.1016/j.brat.2014.04.006

Crichton, F., Dodd, G., Schmid, G., Gamble, G., and Petrie, K. J. (2014). Can expectations produce symptoms from infrasound associated with wind turbines? *Health Psychol.* 33, 360–364. doi: 10.1037/a0031760

Faasse, K., Grey, A., Horne, R., and Petrie, K. J. (2015b). High perceived sensitivity to medicines is associated with higher medical care utilisation, increased symptom reporting and greater information-seeking about medication. *Pharmacoepidemiol. Drug Saf.* 24, 592–599. doi: 10.1002/pds.3751

Faasse, K., Grey, A., Jordan, R., Garland, S., and Petrie, K. J. (2015a). Seeing Is Believing: impact of Social Modeling on Placebo and Nocebo Responding. *Health Psychol.* 34, 880–885. doi: 10.1037/hea0000199

Faasse, K., and Petrie, K. J. (2013). The nocebo effect: patient expectations and medication side effects. *Postgrad. Med. J.* 89, 540–546. doi: 10.1136/postgradmedj-2012-131730

Gaschler, R., Schwager, S., Umbach, V. J., Frensch, P., A., and Schubert, T. (2014). Expectation mismatch: differences between self-generated and cue-induced expectations. *Neurosci. Biobehav. Rev.* 46, 139–157. doi: 10.1016/j.neubiorev.2014.06.009

Hall, K. T., Lembo, A. J., Kirsch, I., Ziogas, D. C., Douaiher, J., and Kaptchuk, T. J. (2012). Catechol-O-methyltransferase val158-met polymorphism predicts placebo effect in irritable bowel syndrome. *PLoS ONE* 7:e48135. doi: 10.1371/journal.pone.0048135

Horne, R., Faasse, K., Cooper, V., Diefenbach, M. A., Leventhal, H., and Petrie, K. J. (2013). The perceived sensitivity to medicines (PSM) scale: an evaluation of validity and reliability. *Br. J. Health Psychol.* 18, 18–30. doi: 10.1111/j.2044-8287.2012.02071.x

Horne, R., and Weinman, J. (2002). Self-regulation and self-management in asthma: exploring the role of illness perceptions and treatment beliefs in explaining non-adherence to preventer medication. *Psychol. Health* 17, 17–32. doi: 10.1080/08870440290001502

Horstmann, G. (2015). The surprise–attention link: a review. *Ann. N. Y. Acad. Sci.* 1339, 106–115. doi: 10.1111/nyas.12679

Kahneman, D., and Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Kaptchuk, T. J., Kelley, J. M., Conboy, L. A., Davis, R. B., Kerr, C. E., and Lembo, A. J. (2008). Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. *Br. Med. J.* 336, 999–1003. doi: 10.1136/bmj.39524.439618.25

Kemper, M., Umbach, V. J., Schwager, S., Gaschler, R., Frensch, P. A., and Sturmer, B. (2012). What I say is what I get: stronger effects of self-generated vs. cue-induced expectations in event-related potentials. *Front. Psychol.* 3:562. doi: 10.3389/fpsyg.2012.00562

Koudriavtseva, T., Onesti, E., Pestalozza, I. F., Sperduti, I., and Jandolo, B. (2012). The importance of physician-patient relationship for improvement of adherence to long-term therapy: data of survey in a cohort of multiple sclerosis patients with mild and moderate disability. *Neurol. Sci.* 33, 575–584. doi: 10.1007/s10072-011-0776-0

Koyama, T., McHaffie, J. G., Laurienti, P. J., and Coghill, R. C. (2005). The subjective experience of pain: where expectations become reality. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12950–12955. doi: 10.1073/pnas.0408576102

Kruschke, J. K. (2003). Attention in learning. *Curr. Dir. Psychol. Sci.* 12, 171–175. doi: 10.1111/1467-8721.01254

Mackintosh, N. J. (1975). Theory of attention - Variations in associability of stimuli with reinforcement. *Psychol. Rev.* 82, 276–298. doi: 10.1037/h0076778

Meissner, K. (2008). Placebo effects on gastric motility and autonomic nervous system activity. *J. Psychosom. Res.* 66, 391–398. doi: 10.1016/j.jpsychores.2008.09.004

Moncrieff, J., Wessely, S., and Hardy, R. (2004). Active placebos versus antidepressants for depression. *Cochrane Database Syst. Rev.* 1, 1–31. doi: 10.1002/14651858.CD003012

Nestoriuc, Y., Orav, E. J., Liang, M. H., Horne, R., and Barsky, A. J. (2010). Prediction of nonspecific side effects in rheumatoid arthritis patients by beliefs about medicines. *Arthritis Care Res.* 62, 791–799. doi: 10.1002/acr.20160

Nestoriuc, Y., von Blanckenburg, P., Schuricht, F., Barsky, A. J., Hadji, P., Albert, U. S., et al. (2016). Is it best to expect the worst? Influence of patients' side-effect expectations on endocrine treatment outcome in a 2-year prospective clinical cohort study. *Ann. Oncol.* 27, 1909–1915. doi: 10.1093/annonc/mdw266

Page, L. A., Petrie, K. J., and Wessely, S. C. (2006). Psychosocial responses to environmental incidents: a review and a proposed typology. *J. Psychosom. Res.* 60, 413–422. doi: 10.1016/j.jpsychores.2005.11.008

Petrie, K. J., Broadbent, E. A., Kley, N., Moss-Morris, R., Horne, R., and Rief, W. (2005). Worries about modernity predict symptom complaints after environmental pesticide spraying. *Psychosom. Med.* 67, 778–782. doi: 10.1097/01.psy.0000181277.48575.a4

Pollo, A., Amanzio, M., Arslanian, A., Casadio, C., Maggi, G., and Benedetti, F. (2001). Response expectancies in placebo analgesia and their clinical relevance. *Pain* 93, 77–84. doi: 10.1016/s0304-3959(01)00296-2

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychol. Rev.* 74, 71–80. doi: 10.1037/h0024109

Rheker, J., Winkler, A., Doering, B. K., and Rief, W. (2016). Learning to experience side effects after antidepressant intake – Results from a randomized, controlled, double-blind study. *Psychopharmacology.* doi: 10.1007/s00213-016-4466-8. [Epub ahead of print].

Rief, W., Barsky, A. J., Bingel, U., Doering, B., Schwarting, R., and Schweiger, U. (2016). Rethinking psychopharmacotherapy: the role of treatment context and brain plasticity in antidepressant and antipsychotic interventions. *Neurosci. Biobehav. Rev.* 60, 51–64. doi: 10.1016/j.neubiorev.2015.11.008

Rief, W., and Glombiewski, J. A. (2012). The hidden effects of blinded, placebo controlled randomized trials: an experimental investigation. *Pain* 153, 2473–2477. doi: 10.1016/j.pain.2012.09.007

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., and Thorwart, A. (2015). Expectations as core features of mental disorders. *Curr. Opin. Psychiatry* 28, 378–385. doi: 10.1097/YCO.0000000000000184

Schedlowski, M., Enck, P., Rief, W., and Bingel, U. (2015). Neuro-bio-behavioral mechanisms of placebo and nocebo responses: implications for clinical trials and clinical practice. *Pharmacol. Rev.* 67, 697–730. doi: 10.1124/pr.114.009423

Schwarz, K. A., Pfister, R., and Büchel, C. (2016). Rethinking explicit expectations: connecting placebos, social cognition, and contextual perception. *Trends Cogn. Sci. (Regul. Ed)* 20, 469–480. doi: 10.1016/j.tics.2016.04.001

Voegtle, E., Barke, A., and Kroener-Herwig, B. (2013). Nocebo hyperalgesia induced by social observational learning. *Pain* 154, 1427–1433. doi: 10.1016/j.pain.2013.04.041

Waszak, F., Cardoso-Leite, P., and Hughes, G. (2012). Action effect anticipation: neurophysiological basis and functional consequences. *Neurosci. Biobehav. Rev.* 36, 943–959. doi: 10.1016/j.neubiorev.2011.11.004

Wendt, L., Albring, A., Benson, S., Engler, H., Engler, A., Hinney, A., et al. (2014). Catechol-O-Methyltransferase Val158Met polymorphism is associated with somatosensory amplification and nocebo responses. *PLoS ONE* 9:e107665. doi: 10.1371/journal.pone.0107665

Winkler, A., and Rief, W. (2015). Effect of placebo conditions on polysomnographic parameters in primary Insomnia: a meta-analysis. *Sleep* 38, 925–932. doi: 10.5665/sleep.4742

Witthöft, M., and Rubin, G. J. (2013). Are media warnings about the adverse health effects of modern life self-fulfilling? An experimental study on idiopathic environmental intolerance attributed to electromagnetic fields (IEI-EMF). *J. Psychosom. Res.* 74, 206–212. doi: 10.1016/j.jpsychores.2012.12.002

Conflict of Interest Statement: 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.

Copyright © 2016 Rief and Petrie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 March 2017
doi: 10.3389/fpsyg.2017.00308

Placebo and Nocebo Effects: The Advantage of Measuring Expectations and Psychological Factors

Nicole Corsi [1, 2] *and Luana Colloca* [1, 3, 4]*

[1] Department of Pain Translational Symptom Science, School of Nursing, University of Maryland, Baltimore, MD, USA, [2] Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy, [3] Department of Anesthesiology/Psychiatry, School of Medicine, University of Maryland, Baltimore, MD, USA, [4] Center to Advance Chronic Pain Research, University of Maryland, Baltimore, MD, USA

Several studies have explored the predictability of placebo and nocebo individual responses by investigating personality factors and expectations of pain decreases and increases. Psychological factors such as optimism, suggestibility, empathy and neuroticism have been linked to placebo effects, while pessimism, anxiety and catastrophizing have been associated to nocebo effects. We aimed to investigate the interplay between psychological factors, expectations of low and high pain and placebo hypoalgesia and nocebo hyperalgesia. We studied 46 healthy participants using a well-validated conditioning paradigm with contact heat thermal stimulations. Visual cues were presented to alert participants about the level of intensity of an upcoming thermal pain. We delivered high, medium and low levels of pain associated with red, yellow and green cues, respectively, during the conditioning phase. During the testing phase, the level of painful stimulations was surreptitiously set at the medium control level with all the three cues to measure placebo and nocebo effects. We found both robust placebo hypolagesic and nocebo hyperalgesic responses that were highly correlated with expectancy of low and high pain. Simple linear regression analyses showed that placebo responses were negatively correlated with anxiety severity and different aspects of fear of pain (e.g., medical pain, severe pain). Nocebo responses were positively correlated with anxiety sensitivity and physiological suggestibility with a trend toward catastrophizing. Step-wise regression analyses indicated that an aggregate score of motivation (value/utility and pressure/tense subscales) and suggestibility (physiological reactivity and persuadability subscales), accounted for the 51% of the variance in the placebo responsiveness. When considered together, anxiety severity, NEO openness-extraversion and depression accounted for the 49.1% of the variance of the nocebo responses. Psychological factors *per se* did not influence expectations. In fact, mediation analyses including expectations, personality factors and placebo and nocebo responses, revealed that expectations were not influenced by personality factors. These findings highlight the potential advantage of considering batteries of personality factors and measurements of expectation in predicting placebo and nocebo effects related to experimental acute pain.

Keywords: acute pain, anxiety, conditioning, expectation, fear, neuroticism, suggestibility

OPEN ACCESS

Edited by:
Mario Gollwitzer, Philipps University of Marburg, Germany

Reviewed by:
Julia Anna Glombiewski, Philipps University of Marburg, Germany
Regine Klinger, Universitätsklinikum Hamburg-Eppendorf (UKE), Germany

***Correspondence:**
Luana Colloca
colloca@son.umaryland.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *08 November 2016*
Accepted: *17 February 2017*
Published: *06 March 2017*

Citation:
Corsi N and Colloca L (2017) Placebo and Nocebo Effects: The Advantage of Measuring Expectations and Psychological Factors. Front. Psychol. 8:308. doi: 10.3389/fpsyg.2017.00308

INTRODUCTION

Personality factors can influence placebo and nocebo effects (Colloca and Grillon, 2014; Colagiuri et al., 2015). Factors such as dispositional optimism (Geers et al., 2005, 2007, 2010; Nes and Segerstrom, 2006; Morton et al., 2009), hypnotic suggestibility (De Pascalis et al., 2002), somatic focus (Geers et al., 2006; Johnston et al., 2012), empathy (Colloca and Benedetti, 2009; Hunter et al., 2014; Rütgen et al., 2015a,b), neuroticism (Peciña et al., 2013), altruism (Peciña et al., 2013), social desirability (Gelfland et al., 1965), dopamine-related traits (Schweinhardt et al., 2009), fear of pain (Flaten et al., 2006; Zubieta et al., 2006; Lyby et al., 2010), locus of ego-resilience (Peciña et al., 2013), anxiety (Staats et al., 2001; Ober et al., 2012), pessimism (Geers et al., 2005; Corsi et al., 2016), pain catastrophizing (Vogtle et al., 2013), harm avoidance, and persistence (Corsi et al., 2016) have been linked to placebo and nocebo effects.

In particular, optimism, the active behavioral and mental coping ability of individuals to face adversity, has been liked to proneness to show higher placebo analgesic effects (Geers et al., 2005, 2007, 2010). Attention toward the body, referred as somatic focus, is related to larger placebo analgesic effects and higher positive expectations (Geers et al., 2006). Empathic resonance and concern for others have been linked to placebo analgesia as well (Colloca and Benedetti, 2009; Hunter et al., 2014; Rütgen et al., 2015a,b). Hypnotic susceptibility and responsiveness to verbal suggestions influence placebo analgesia (Huber et al., 2013). Other factors such as Neuroticism-Extraversion-Openness to experience (NEO), NEO Altruism, NEO Straightforwardness, NEO Angry Hostility and Ego-Resiliency, have been coupled with a 25% variance in behavioral placebo responses to pain and 27% of the μ-opioid system activation in the nucleus accumbens (Peciña et al., 2013).

Conversely, anxiety (Staats et al., 2001), harm avoidance and persistence (Corsi et al., 2016) and pain catastrophizing (Swider and Babel, 2013; Vogtle et al., 2013) have been associated with nocebo effects. Anxiety and harm avoidance correlate positively with nocebo effects, while optimism and persistence correlate negatively with nocebo effects in the context of the motor system (Corsi et al., 2016). In the present study, our aim was to investigate how distinct positive and negative personality factors estimate the likelihood of placebo and nocebo effects. Moreover, we aimed to establish the relationship among trial-by-trial expectations of pain reduction and increase, and placebo/nocebo effects, and personality. We hypothesized that using aggregated personality factors and expectations would allow us to better estimate placebo and nocebo responses in a laboratory setting using a well-established conditioning model (Colloca et al., 2010).

MATERIALS AND METHODS

Study Participants

We recruited 50 participants from Baltimore, MD, USA to enroll a total of 46 healthy participants (24 women; 27.41 ± 1.07 years; see **Table 1**). Four participants were excluded: two of them did not meet the inclusion criteria and two were unable to discriminate distinct levels of heat thermal stimulation that are used for the acquisition phase of the conditioning paradigm. Upon arrival, participants signed a consent form to study pain modulation. Participants with cardiovascular and neurological diseases, family or personal history of psychiatric conditions, personal history of drug abuse, acute or chronic pain, color blindness, impaired hearing, pregnancy and current use of painkillers and any other medication, were excluded from participating in this study. On the day of the experiment, a toxicology drug test was also performed to exclude any recent use of marijuana, cocaine, opiates such as hydrocodone, oxycodone and hydromorphone, amphetamine, methamphetamine, ecstasy/MDMA and phencyclidine. Participants who reported use of tobacco or nicotine over the last year were also excluded.

TABLE 1 | Characteristics of study participants.

Characteristics of Participants	
Sex	24 females
	22 males
Age (years)	27.41 ± 1.07
Body Mass Index (BMI)	26.00 ± 0.71
Systolic blood pressure values (mmHg)	120.19 ± 2.00
Diastolic blood pressure (mmHg)	75.15 ± 1.27
Heart rate (beats per minute)	66.36 ± 1.45
Levels of pain (°C)	Low 41.51 ± 0.36
	Medium 44.55 ± 0.36
	High 47.52 ± 0.36

All values are expressed as mean ± SE.

This study was carried out in accordance with the recommendations of the UMB Institutional Review Board with written informed consent from all subjects.

All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the UMB Ethics Committee (Prot # HP00065783). Due to the use of deception, a debriefing written form was given to each participant at the end of the study participation offering to withdraw the data from the study. None of them opted to do so. Participants were compensated for their participation ($90).

Pain Assessment

A well-validated paradigm that has been previously described (Colloca et al., 2010) was used to explore placebo and nocebo responses to a contact heat thermal painful stimulation.

Individual pain sensitivity and tolerance were measured in each participant using the ATS Medoc Pathway system (Medoc Advanced Medical System, Rimat Yishai, Israel). A 3 × 3 cm thermode was placed on the dominant forearm as confirmed by the Edinburgh Handedness Inventory. The baseline temperature delivered by the Medoc equipment was 32°C. Ascending series of stimulations starting from warm sensation to maximum tolerable pain were delivered, while the participant was asked to stop the machine as soon as she felt a warm sensation, low, medium and high pain. Each level was assessed four times and averaged to determine the intensities of stimulations to be used during the acquisition and testing phases of the conditioning paradigm. We defined then the painful stimulations by subtracting 3 and 6°C

starting from the highest reported level of tolerable pain (e.g., 49 and 43°C) so that the levels of stimulation were standardized among participants. The intensities of stimulation were also rated to ensure correspondence to individual experience of low, medium and high pain.

Placebo and Nocebo Manipulation

Three visual cues (red, yellow, and green) were displayed on a computer placed one meter apart from a chair in a quiet lab. Participants were told that the green, yellow and red lights would anticipate the delivery of a low, medium and high level of pain, respectively.

During the acquisition phase of the classical conditioning paradigm, 18 painful stimulations were delivered at the three levels of pain corresponding to an individual low, medium, and high level of pain in association to six red, six yellow, and six green cues, respectively. Afterwards, during the testing phase, 9 stimulations were paired with the three color cues but the intensity was set at same medium control level in accordance with a previously described paradigm (Colloca et al., 2010). The sequence of the cue presentation was counterbalanced across participants using four distinct sequences. This change in the pain levels allowed us to explore how first-hand experience of low and high pain during the acquisition phase results in placebo and nocebo responses during the testing phase. Participants rated the experienced pain immediately after the painful stimulation using the VAS scale (from 0 = no pain to 100 = maximum tolerable pain). Pain reports were collected using Celeritas Fiber Optic Response System (Psychology Software Tools, Inc., Sharpsburg, PA, USA).

Moreover, trial-by-trial expectations were measured. The terms "expectation" and "expectancy" have been often used in an interchangeable way. Herein, we adopted the term "expectation" to refer to verbalized and measurable constructs as compared to "expectancies" defining psychophysical predictions that can be present without full awareness (i.e., implicit expectancies) (Kube et al., 2017).

Participants were asked to rate their expectations of the upcoming stimulation immediately before the delivery of the thermal stimulation using a VAS anchored from 0 = no pain to 100 = maximum tolerable pain.

During each trial, the visual cue was presented for 4 s. Immediately after the presentation of the cue, participants were asked to rate their expectation (5 s) about the upcoming stimulus. The thermal stimulation lasted for 10 s. Then participants were asked to rate their perceived pain (5 s) and an inter-trial interval followed with a variable timing (8–10 s). The procedure and the delivery of painful stimulations were controlled by scripts pre-programmed in Eprime (Psychology Software Tools, Inc., Sharpsburg, PA, USA; version 2.0). To prevent habituation, the presentation of visual cues during both phases was counterbalanced using four preprogramed sequences.

Psychological Questionnaires

Participants completed a comprehensive battery of psychological questionnaires, which were chosen to cover distinct psychological factors that we hypothesized to be linked to placebo and nocebo effects. In particular, for the placebo-related factors, we included optimism, reward, suggestibility, empathy and sensation-seeking and motivation. We used the following questionnaires: (1) Life-Orientation Test-Revisited, Lot-R (Scheier et al., 1994) to assess generalized optimism vs. pessimism; (2) Behavioral Inhibition and Behavioral Activation Scale, BIS/BAS (Carver and White, 1994) to investigate dispositional sensitivity to the behavioral inhibition system (BIS) and the behavioral activation system (BAS); (3) Multidimensional Iowa Suggestibility Scale, MISS (Kotov et al., 2004) to investigate the main components of suggestibility; (4) Interpersonal Reactivity Index, IRI (Davis, 1980) to measure the participant's dispositional empathy in different situations; (5) Sensation Seeking (SS) (Zuckerman, 1994) to measure the necessity to find and experience new situations; (6) Tri-dimensional Personality Questionnaire, TPQ (Cloninger et al., 1991) to assess novelty seeking (NS), harm avoidance (HA), and reward dependence (RD); (7) and the Intrinsic Motivation Inventory (IMI) (Markland and Hardy, 1997) to assess participants' experience during the experimental procedure that was just performed.

For the nocebo-related psychological factors included measurements of various aspects of anxiety (e.g., state, severity, and sensitivity), catastrophizing, neuroticism, fear of pain, depression and feelings of worry. The following inventories were used: (1) State and Trait Anxiety Inventory, STAI (Spielberger, 1983) to investigate anxiety either in a precise moment (STAI-Y1) or as a general tendency (STAI-Y2); (2) Anxiety Sensitivity Index, ASI (Reiss et al., 1986) to assessed beliefs of sensations that could have harmful consequences; (3) Beck Anxiety Inventory, BAI (Beck et al., 1988) to measure experience of anxiety symptoms during the previous 2 weeks; (4) Beck Depression Inventory, BDI (Beck et al., 1961) to include items relating to depression, cognitions, as well as physical symptoms; (5) Mood and Anxiety Symptom Questionnaire, MASQ (Haigh et al., 2011) to assess depressive symptoms and anxiety symptoms; (6) Pain Catastrophizing Scale, PCS (Sullivan et al., 1995) to assess catastrophizing impacts on pain experience; (7) Neuroticism—Extroversion—Openness Inventory (NEO)—Five Factory Inventory (FFI) (Costa and McCrae, 1985, 1992) to investigate Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness; (8) Fear of Pain Questionnaire, FOP (Osman et al., 2002) to measure fear levels to different types of physical pain; (9) Penn State Worry Questionnaire, PSWQ (Meyer et al., 1990) to measure the trait of worry in different situations.

We also administered the Positive and Negative Affective Schedule, PANAS (Crawford and Henry, 2004), that investigates the relationships between positive and negative affect with personality states and emotions.

Statistical Analysis

VAS pain and VAS expectations ratings were compared using repeated measure ANOVA. We tested for the main effect of the factor condition (red, yellow, and green) and time (trials) set both as within-subjects factors. *F*-tests were followed by the Bonferroni *post-hoc* tests for multiple comparisons. We also

tested for sex influences on placebo and nocebo effects using sex as a between factor. Partial eta squared (η^2) effect sizes are reported for all the comparisons.

VAS pain and expectation scores from the testing phase were further averaged across trials to calculate the difference between yellow-green and yellow-red pain scores to be correlated with placebo and nocebo effects, respectively.

The above psychological questionnaire scores were used in both simple correlation and multivariate analyses. We analyzed psychological questionnaire scores using both Spearman correlation and stepwise multiple regression model analyses in which the questionnaires were modeled to predict placebo and nocebo responses. Mediation analyses were also calculated with expectation as mediator (M), placebo (or nocebo) responses as dependent variable (Y), and personality factors as independent variable (X). For testing indirect effects, a bootstrapping method based on resampling of 1,000 times was used in accordance with Preacher and Hayes methods (Preacher and Hayes, 2004; Hayes and Preacher, 2010). All the analyses were carried out using the SPSS software package (SSPS Inc, Chicago, Illinois, USA, vers.21). To minimize alpha errors, the level of significance was set at $p \leq 0.005$.

RESULTS

We performed separate analyses for the VAS pain and expectation reports related to the acquisition and testing phases of the conditioning paradigm.

Conditioning: Acquisition Phase

We analyzed the VAS pain reports during the acquisition phase, and found that participants distinguished the low, medium and high levels of painful stimuli [main effect of condition: $F_{(2, 88)} = 503.970$, $p < 0.001$, $\eta^2 = 0.920$]. The average pain score for red-associated stimuli was 74.73 ± 2.36 using an average intensity of pain equal to 47.52°C, the average pain score for yellow was 29.55 ± 1.54 using an average pain equal to 44.55°C and the average pain score for green was 9.37 ± 0.96 when an average pain equal to 41.51°C out of 50°C was delivered. The factor time was significant [$F_{(5, 220)} = 7.359$, $p < 0.001$, $\eta^2 = 0.143$]. The condition × time interaction was significant [$F_{(10, 440)} = 5.324$, $p < 0.001$, $\eta^2 = 0.108$] (**Figure 1A**) showing a quadratic trajectory [$F_{(1, 44)} = 10.308$, $p < 0.002$, $\eta^2 = 0.190$].

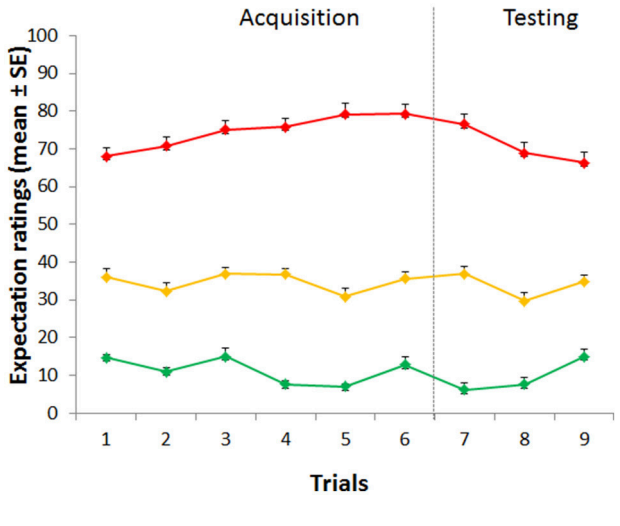


FIGURE 2 | Time course of expectation ratings. Expectations during the acquisition phase differed across the three conditions. During the testing phase, expectations for high, medium and low pain continued to be staidly different across the three conditions.

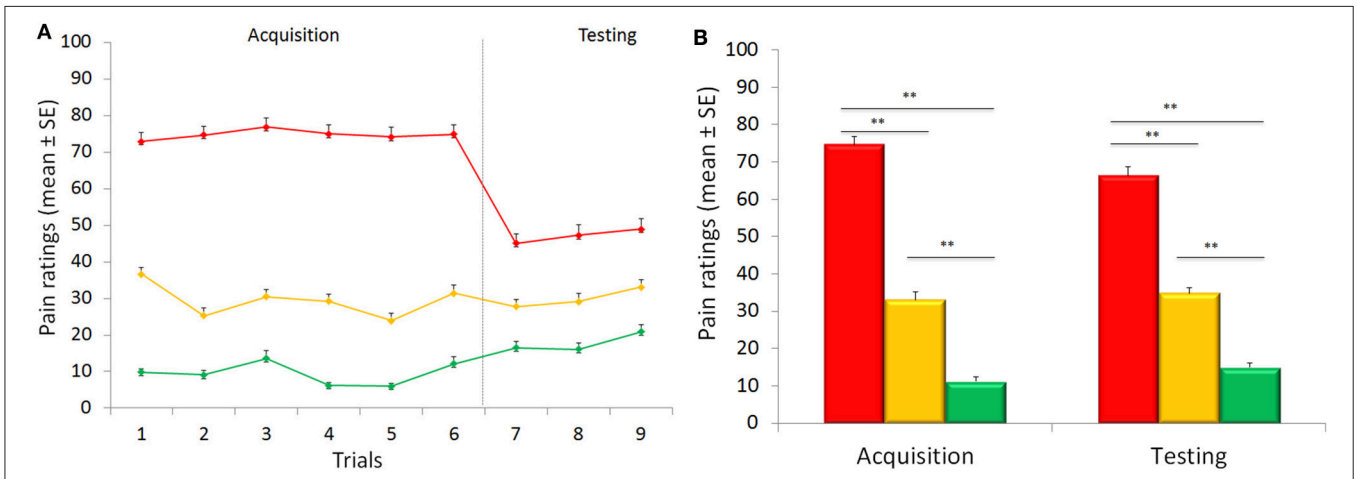


FIGURE 1 | Time course of placebo and nocebo responses (A). Representation trial-by-trial of the average of pain ratings for control (yellow), placebo (green) and nocebo (red) responses during the acquisition (trials 1–6) and the testing (trials 7–9) phases. Participants learned to distinguish the low, medium and high levels of painful stimuli over the acquisition phase. During the testing phase, there was a significant placebo and nocebo effect indicating no extinction over the entire experimental session. Graphical representation of the pain score for the red, green, and yellow associated stimuli **(B)**. The red associated stimuli were perceived as higher than the yellow control stimuli and green were rated as lower than the yellow stimuli during the testing phases when the stimulation was surreptitiously set at a medium level for the three colors indicating both robust placebo and nocebo effects. Data are expressed as mean ± sem. $^{**}p < 0.001$.

VAS expectation scores (75.63 ± 2.09, 34.74 ± 1.61, and 11.30 ± 0.98, respectively) during the acquisition phase differed across the three conditions [$F_{(2, 88)} = 515.152$, $p < 0.001$; $\eta^2 = 0.921$], with significant time [$F_{(5, 220)} = 3.392$, $p = 0.006$; $\eta^2 = 0.072$] and condition × time interaction [$F_{(10, 440)} = 7.542$, $p < 0.001$; $\eta^2 = 0.146$] effects (**Figure 2**).

Conditioning: Testing Phase

During the testing phase, when the level of pain was set at the same control (yellow) intensity for the three cues, VAS pain reports revealed a significant effect of condition [$F_{(2, 88)} = 96.04$, $p < 0.001$; $\eta^2 = 0.686$], time [$F_{(2, 88)} = 7.553$, $p = 0.001$; $\eta^2 = 0.147$] with a non-significant condition × time interaction [$F_{(4, 176)} = 0.378$, $p = 0.824$; $\eta^2 = 0.009$] indicating no extinction over the entire experimental session (**Figure 1A**). *Post-hoc* Bonferroni tests indicated that the red stimuli (average VAS: 46.98 ± 2.46) were perceived as higher than the yellow control stimuli (average VAS: 29.96 ± 1.78) ($p < 0.001$) and green (average VAS: 17.86 ± 1.70) were rated as lower than the yellow stimuli ($p < 0.001$) indicating both robust placebo and nocebo effects (**Figure 1B**). The distribution and magnitude of placebo and nocebo responses ranged from no effects to large changes in pain modulation (**Figure 3**).

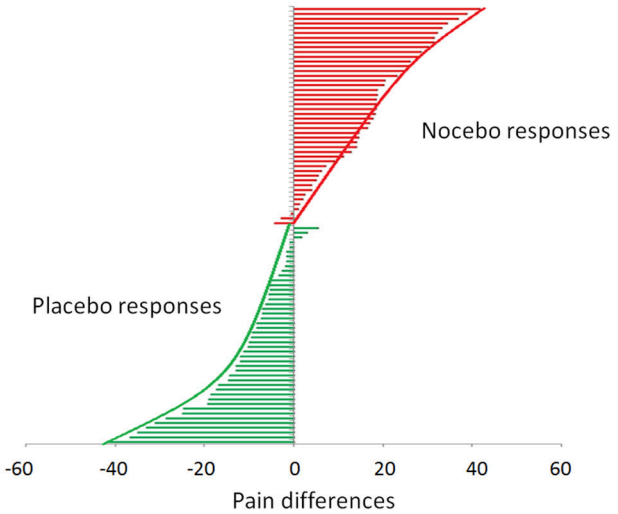


FIGURE 3 | Distribution of placebo and nocebo effects. Each bar represents a single study participant. The green bars represent the magnitude of the placebo effect (yellow-green VAS scores). The red bars represent the magnitude of the nocebo effect (yellow-red-VAS score). It is worth noting that the individual placebo and nocebo responses range from no responses at all to medium to large effect.

Placebo effects were significantly correlated with the hypoalgesic effect experienced during the acquisition phase (Placebo: $r = 0.388$, $p = 0.008$) but nocebo hyperalgesic responses appeared to be independent of the experienced high pain ($r = 0.080$, $p = 0.598$). Moreover, being prone to experience a placebo response did not imply being also prone to experience a nocebo response, as indicated by the absence of significant correlation between individual placebo and the nocebo responses ($r = -0.113$, $p = 0.454$).

During the testing phase, expectations for high, medium and low pain [70.61 ± 2.45, 33.87 ± 1.81, and 9.54 ± 0.93] were different across the three conditions [$F_{(2, 88)} = 441.355$, $p < 0.001$; $\eta^2 = 0.909$] with a main effect of time [$F_{(2, 88)} = 8.092$, $p = 0.001$; $\eta^2 = 0.155$], and a significant interaction condition × time [$F_{(4, 176)} = 13.156$, $p < 0.001$; $\eta^2 = 0.230$] (**Figure 3**), showing a linear trajectory [$F_{(1, 44)} = 33.850$, $p < 0.001$, $\eta^2 = 0.435$]. Importantly, we found that positive expectations correlated with placebo responses ($r = 0.412$, p = 0.002, **Figure 4A**) and similarly negative expectations correlated with nocebo effects ($r = 0.351$, $p = 0.008$, **Figure 4B**).

In this cohort of participants, sex effects for placebo, nocebo and expectancies were not observed [placebo: $F_{(1, 44)} = 0.010$, $p = 0.922$; nocebo: $F_{(1, 44)} = 0.990$, $p = 0.325$; positive

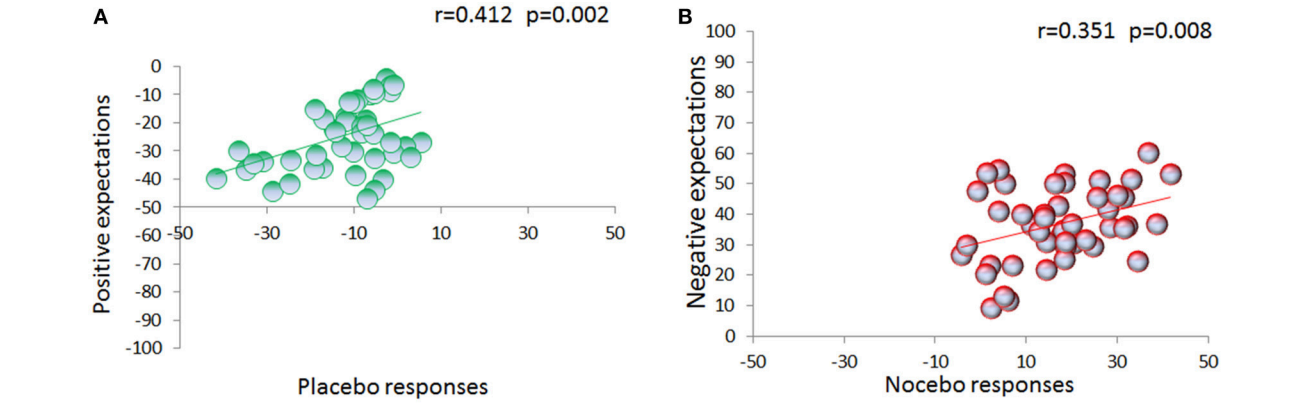


FIGURE 4 | Relation between expectations and placebo/nocebo effects. VAS expectation scores were collected on a trial-by-trial basis during the testing phase. Expectation of low pain positively correlates with placebo effects **(A)**. Similarly, expectation of upcoming high painful stimulation positively correlates with nocebo effects **(B)**.

TABLE 2 | Correlations between placebo, nocebo and personality factors.

Personality factors	PLACEBO EFFECT		NOCEBO EFFECT	
	R	*p*	*r*	*p*
STAI-Y1 pre	−0.175	0.246	0.217	0.147
STAI-Y2 post	−0.004	0.978	0.177	0.241
STAY2	−0.123	0.415	0.119	0.432
ASI	−0.147	0.330	**0.460**	**0.001**
BAI	**−0.485**	**0.001**	−0.028	0.855
BDI	−0.039	0.796	0.244	0.102
PANAS total	0.096	0.527	0.175	0.245
PANAS positive	−0.218	0.145	−0.199	0.429
PANAS negative	−0.366	0.012	0.248	0.097
NEO neuroticism	−0.166	0.270	0.021	0.892
NEO extraversion	−0.349	0.018	0.186	0.217
NEO openness	−0.228	0.128	−0.264	0.076
NEO agreeableness	−0.098	0.518	0.091	0.548
NEO conscientiousness	−0.001	0.993	0.054	0.720
MASQ General depressive scale	−0.75	0.622	−0.051	0.735
MASQ anxious arousal	0.021	0.889	−0.098	0.519
MASQ general distress	−0.146	0.332	0.155	0.305
MASQ anhedonia	0.015	0.620	0.151	0.316
TPQ novelty seeking	0.015	0.922	0.045	0.769
TPQ harm avoidance	−0.164	0.277	0.099	0.511
TPQ reward dependence	−0.191	0.204	0.179	0.234
BAS drive	0.165	0.273	−0.016	0.918
BAS fun	0.000	0.999	−0.139	0.356
BAS reward	0.004	0.977	−0.32	0.382
BIS	−0.065	0.670	−0.091	0.549
BIS/BAS total	0.013	0.934	−0.144	0.341
LotR	−0.072	0.635	0.012	0.935
IMI interest/enjoyment	−0.075	0.620	−0.081	0.595
IMI perceived competence	−0.280	0.060	−0.022	0.883
IMI effort/importance	−0.142	0.346	−0.220	0.143
IMI pressure/tense	−0.017	0.910	0.295	0.047
IMI choice	−0.094	0.540	−0.197	0.194
IMI value/utility	−0.343	0.020	−0.216	0.149
IMI total	−0.339	0.021	−0.144	0.341
IRI fantasy	−0.290	0.050	−0.016	0.915
IRI empathic concern	−0.231	0.123	0.157	0.297
IRI perspective-taking	−0.189	0.208	0.093	0.538
IRI personal distress	0.190	0.207	0.195	0.193
MISS suggestibility	−0.165	0.274	0.254	0.089
MISS persuadability	−0.060	0.693	−0.016	0.918
MISS physiological suggestibility	−0.264	0.076	**0.438**	**0.002**
MISS physiological reactivity	−0.354	0.016	0.159	0.292
MISS peer conformity	−0.270	0.069	0.293	0.048
MISS mental control	−0.220	0.141	0.005	0.975
MISS unpersuadability	−0.121	0.421	0.295	0.047
MISS short suggestibility	−0.284	0.056	0.175	0.245
MISS total	−0.331	0.025	0.301	0.042
FOP severe	**−0.490**	**0.001**	−0.073	0.629
FOP medical	**−0.416**	**0.004**	−0.013	0.929
FOP total	**−0.435**	**0.003**	−0.037	0.806
PCS rumination	−0.104	0.490	0.352	0.016
PCS magnification	0.032	0.831	0.054	0.721
PCS helplessness	0.021	0.887	0.366	0.012
PCS total	−0.022	0.883	0.343	0.020
PSWQ	−0.216	0.149	0.283	0.057
SS boredom susceptibility	−0.066	0.661	0.098	0.518
SS disinhibition	−0.031	0.839	−0.027	0.861
SS experience seeking	0.078	0.605	−0.093	0.537
SS adventure seeking	−0.036	0.812	−0.025	0.869
SS total	0.014	0.924	0.005	0.976

STAI 1-2, State and Trait Anxiety Inventory; ASI, Anxiety Sensitivity Index; BAI, Beck Anxiety Inventory; BDI, Beck Depression Inventory; PANAS, Positive and Negative Affective Schedule; NEO, Neuroticism-Extraversion-Openness Inventory; MASQ, Mood and Anxiety Symptom Questionnaire; TPQ, Tridimensional Personality Questionnaire; BISBAS, Behavioral Inhibition and Behavioral Activation Scale; LotR, Life-Orientation Test-Revisited; IMI, Intrinsic Motivation Inventory; IRI, Interpersonal Reactivity Index; MISS, Multidimensional Iowa Suggestibility Scale; FOP, Fear of Pain; PCS, Pain Catastrophizing Scale; PSWQ, Penn State Worry Questionnaire; SS, Sensation Seeking. Significant results are indicated in bold.

expectancies: $F_{(1, 44)} = 1.860$, $p = 0.180$; negative expectancies: $F_{(1, 44)} = 0.025$, $p = 0.875$].

Personality Predictors

We then explored the effects of personality factors on placebo and nocebo effects. First, we ran a series of correlations analyses and found that placebo responses were negatively correlated with severity of anxiety (BAI: $r = -0.485$, $p = 0.001$), and fear of pain (FOP, severe: $r = -0.490$, $p = 0.001$; medical fear, $r = -0.416$, $p = 0.004$; total fear $r = -0.435$, $p = 0.003$). By the contrary, nocebo responses were positively correlated with anxiety sensitivity (ASI, $r = 0.460$, $p = 0.001$), physiological suggestibility (MISS: $r = 0.438$, $p = 0.002$) with a trend for catastrophizing tendency (PCS rumination: $r = 0.352$, $p = 0.016$; PCS helplessness: $r = 0.366$, $p = 0.012$; PCS total: $r = 0.343$, $p = 0.020$) (**Table 2**).

Moreover, we considered the hypothesized psychological factors taken together in order to identify their relationship with the dependent variables (e.g., placebo and nocebo VAS) using stepwise multiple regression models. The significant values are reported in **Tables 3, 4**. Motivation (value/utility and pressure/tense subscales) and suggestibility (physiological reactivity and persuadability subscales) accounted for 51% of variance in placebo responses (**Table 3**). Conversely, ASI, NEO-openness-extraversion and depression taken together accounted for 49.1% of variance in nocebo responses (**Table 4**).

Finally, we calculated mediation analyses for exploring the relationship among personality factors, positive/negative expectations and placebo/nocebo responses. Interestingly, we found that expectations were significantly linked to placebo and nocebo effects (see **Table 5**). However, personality factors *per se* did not influence expectancies, and the indirect effect among the three variables was not significant. Due to the exploratory

TABLE 3 | Stepwise multiple regression models for the prediction of placebo effects.

Dependent variable	Predictor Variables	R^2	β	*t*	*p*
Placebo hypoalgesia	*Model 1*	21.6			
	MISS physiol		0.464	3.438	0.001
Placebo hypoalgesia	*Model 1*	21.6			
	MISS physiol		0.464	3.438	0.001
	Model 2	35.4			
	MISS physiol		0.577	4.452	< 0.001
	IMI value		−0.389	−2.999	0.005
Placebo hypoalgesia	*Model 1*	21.6			
	MISS physiol		0.464	3.438	0.001
	Model 2	35.4			
	MISS physiol		0.577	4.452	< 0.001
	IMI value		−0.389	−2.999	0.005
	Model 3	42.7			
	MISS physiol		0.579	4.687	< 0.001
	IMI value		−0.371	−2.993	0.005
	MISS persuadability		0.270	2.280	0.028
Placebo hypoalgesia	*Model 1*	21.6			
	MISS physiol		0.464	3.438	0.001
	Model 2	35.4			
	MISS physiol		0.577	4.452	< 0.001
	IMI value		−0.389	−2.999	0.005
	Model 3	42.7			
	MISS physiol		0.579	4.687	< 0.001
	IMI value		−0.371	−2.993	0.005
	MISS persuadability		0.270	2.280	0.028
	Model 4	51.0			
	MISS physiol		0.463	3.745	0.001
	IMI value		−0.335	−2.871	0.007
	MISS persuadability		0.344	3.006	0.005
	IMI pressure		0.319	2.617	0.012

MISS, Multidimensional Iowa Suggestibility Scale (Physiological Reactivity and Persuadability subscales); IMI, Intrinsic Motivation Inventory (Value/Utility and Pressure/Tense subscales). Only significant values are shown. Excluded variables (not significant): Lot-R, Life-Orientation Test-Revisited; BIS/BAS, Behavioral Inhibition and Behavioral Activation Scale; IRI, Interpersonal Reactivity Index; SS, Sensation Seeking; TPQ, Tri-dimensional Personality Questionnaire.

TABLE 4 | Stepwise multiple regression models for the prediction of nocebo effects.

Dependent variable	Predictor Variables	R^2	β	*t*	*p*
Nocebo hyperalgesia	*Model 1*	20.3			
	ASI		0.451	3.349	0.002
Nocebo hyperalgesia	*Model 1*	20.3			
	ASI		0.451	3.349	0.002
	Model 2	33.3			
	ASI		0.498	3.966	< 0.001
	NEO_O		−0.364	−2.897	0.006
Nocebo hyperalgesia	*Model 1*	20.3			0.002
	ASI		0.451	3.349	
	Model 2	33.3			
	ASI		0.498	3.966	< 0.001
	NEO openess		−0.364	−2.897	0.006
	Model 3	42.9			
	ASI		0.493	4.197	< 0.001
	NEO openess		−0.472	−3.796	< 0.001
	NEO extraversion		0.329	2.660	0.011
Nocebo hyperalgesia	*Model 1*	20.3			
	ASI		0.451	3.349	0.002
	Model 2	33.3			
	ASI		0.498	3.966	< 0.001
	NEO openess		−0.364	−2.897	0.006
	Model 3	42.9			
	ASI		0.493	4.197	< 0.001
	NEO openess		−0.472	−3.796	< 0.001
	NEO extraversion		0.329	2.660	0.011
	Model 4	49.1			
	ASI		0.448	3.919	< 0.001
	NEO openess		−0.413	−3.388	0.002
	NEO extraversion		0.387	3.197	0.003
	BDI		0.267	2.218	0.032

ASI, Anxiety Sensitivity Index; BDI, Beck Depression Inventory; NEO, Neuroticism-Extraversion-Openness Inventory. Only significant values are shown. Excluded variables (not significant for the model): STAI, State and Trait Anxiety Inventory; BAI, Beck Anxiety Inventory; FOP. Fear of Pain; MASQ, Mood and Anxiety Symptom Questionnaire; PCS, Pain Catastrophizing Scale; PSWQ, Penn State Worry Questionnaire; PANAS, Positive and Negative Affective Schedule.

nature of this part of the study, we used a relative broad battery. Therefore, correlations among personality questionnaires are shown in **Table 6**.

DISCUSSION

In this study, we investigated the influence of expectations and hypothesized psychological factors on placebo and nocebo effects elicited by a well-established model of conditioning and heat thermal painful stimulation. Placebo hypoalgesic responses were negatively correlated with severity of anxiety and fear of pain (e.g., medical fear, severe, and total fear). On the contrary, nocebo hyperalgesic responses were positively correlated with anxiety sensitivity, suggestibility and catastrophizing (trend only). Moreover, a stepwise regression modeling showed that aggregate scores of Motivation (value/utility and pressure/tense subscales) and suggestibility (physiological reactivity and persuadability subscales) accounted for the 51% of the variance in the placebo responses. By contrast, the aggregation of anxiety, openness, extraversion and depression accounted for the 49.1% of the variance in the nocebo responses. Importantly, expectations were highly correlated with placebo and nocebo effects and psychological factors did not influence level of expectations towards reduction or increase of pain.

Consistently with previous studies (Colloca and Benedetti, 2006, 2009; Colloca et al., 2008, 2010; Lui et al., 2010), we found

TABLE 5 | Mediation analysis results.

Model	a path	b path	c' path	Indirect effect
BAI (X) Positive expectations (M) Placebo hypoalgesia (Y)	$p = 0.094$	$p = 0.018$	$p = 0.099$	$p = 0.184$
FOP severe (X) Positive expectations (M) Placebo hypoalgesia (Y)	$p = 0.656$	$p = 0.005$	$p = 0.005$	$p = 0.673$
FO*P* medical (X) Positive expectations (M) Placebo hypoalgesia (Y)	$p = 0.656$	$p = 0.011$	$p = 0.012$	$p = 0.309$
FOP total (X) Positive expectations (M) Placebo hypoalgesia (Y)	$p = 0.217$	$p = 0.007$	$p = 0.012$	$p = 0.286$
MISS physiol. (X) Negative expectations (M) Nocebo hyperalgesia (Y)	$p = 0.181$	$p = 0.010$	$p = 0.078$	$p = 0.321$
ASI (X) Negative expectations (M) Nocebo hyperalgesia (Y)	$p = 0.871$	$p < 0.001$	$p = 0.014$	$p = 0.879$
PCS rumination (X) Negative expectations (M) Nocebo hyperalgesia (Y)	$p = 0.493$	$p = 0.023$	$p = 0.006$	$p = 0.539$
PCS help. (X) Negative expectations (M) Nocebo hyperalgesia (Y)	$p = 0.322$	$p = 0.031$	$p = 0.022$	$p = 0.398$
PCS total (X) Negative expectations (M) Nocebo hyperalgesia (Y)	$p = 0.350$	$p = 0.027$	$p = 0.014$	$p = 0.419$

BAI, Beck Anxiety Inventory; FOP, Fear Of Pain; MISS, Multidimensional Iowa Suggestibility Scale; ASI, Anxiety Sensitivity Index; PCS, Pain Catastrophizing Scale.

that visual cues associated with prior experiences of low and high pain elicit strong placebo and nocebo effects with a distribution raging from no responses to low modulation of pain, to medium and high reductions and increases (**Figure 3**). Studies on placebo hypoalgesia and nocebo hyperalgesia have shown a substantial inter-individual variability and distinct personality factors have been associated with placebo and nocebo effects (Colloca and Grillon, 2014; Colagiuri et al., 2015). There is evidence that some personality factors such as anxiety (Staats et al., 2001; Ober et al., 2012), fear of pain (Lyby et al., 2010) and neuroticism (Peciña et al., 2013), are associated with reduced placebo analgesia. We confirmed and expanded some of these findings. In our study, severity of anxiety as well as fear of pain (e.g., medical, sever, and total fear) were linked to reduced placebo responsiveness to pain. Severity of anxiety including symptoms of depression, feelings of hopelessness and irritability, guiltiness or feelings of being punished, as well as physical symptoms such as fatigue, correlated negatively with placebo responses with higher severity of anxiety linked to lower reduction of pain induced by positive expectations. High levels of fear of pain referring to the dispositional tendency to have negative emotions toward pain and pain anticipation have been also associated with placebo- and nocebo-induced pain modulation (Lyby et al., 2010; Aslaksen and Lyby, 2015). We found that fear of medical pain in particular correlates with low placebo hypoalgesic responses and this is consistent with the parallel enhancement of nocebo induced by fear of pain and other medical procedures (Aslaksen and Lyby, 2015).

When we looked at the nocebo effect—the negative counterpart of the placebo phenomenon (Petrovic, 2008)—we found a positive correlation with anxiety sensitivity, physiological suggestibility and catastrophizing. Anxiety sensitivity refers to behaviors or sensations associated with the experience of anxiety that elicit misinterpretations of bodily sensations such as the experience of a no harmful stimulus causing intense pain (Mehta et al., 2016). Suggestibility is a trait-like characteristic creating distinct behaviors that facilitate responsiveness to plausible information as well as inclinations to accept and act on others' suggestions in regards to the body (e.g., physical suggestibility), and has been linked to placebo effects (Lund et al., 2015) and nocebo effects (Corsi et al., 2016). Catastrophizing, a maladaptive cognitive process that is potentially heritable and has been reported to predict severity of clinical pain (Flor and Turk, 1988; Severeijns et al., 2001; Goubert et al., 2004; Kudel et al., 2005; Trost et al., 2015), has been recently explored and shown to be relevant for nocebo effects (Vogtle et al., 2013).

Personality is a continuum of factors and thus highlights the importance of considering distinct factors together. Therefore, based on the literature we took into consideration two sets of psychological factors related to placebo and nocebo responsiveness and used a multilevel modeling approach in which hierarchies and residual components at each level within a hierarchy are computed. Such an approach indicated that an aggregate score for motivation (value/utility and pressure/thanks subscales) and suggestibility (physiological reactivity and persuadability subscales) accounted for the 51% of the variance in the placebo hypolagesic responses whilst anxiety severity, NEO-openness-extraversion and depression considered together accounted for the 49.1% of the variance of nocebo responses suggesting that it helps evaluate the psychological factors comprehensively. Another important result from this study was that positive expectations were significantly correlated with placebo responses and negative expectations were significantly correlated with nocebo responses. Although one may argue that asking on a trial-by-trial about expectancy of the upcoming pain may have generated a sort of self-prophecy (e.g., You get what you expect, you get what you ask for), it remains an interesting finding that could be important to keep in mind every time we measure pain in real-world settings. Therefore, an obvious question was whether personality factors impact the formation of expectations of pain reduction and increase. In this study, mediation analyses indicated that personality factors (e.g., being worried, being fearful) had no direct effect on the

TABLE 6 | Correlations among personality measurement tools.

		STAI post	STAI 2	ASI	BAI	BDI	PANAS	NEO	MASQ	TPQ	BIS/BAS	LotR	IMI	IRI	MISS	FOPtot	PCS	PSWQ	SS
STAI_1 pre	Rho	0.690**	0.689**	0.277	0.153	0.336*	0.351*	0.029	0.607**	0.175	−0.054	−0.147	−0.041	0.057	0.072	0.213	0.214	0.170	0.020
	p-value	**0.000**	**0.000**	0.062	0.309	0.022	0.017	0.850	**0.000**	0.246	0.722	0.328	0.789	0.709	0.636	0.155	0.154	0.258	0.898
STAI_1 post	rho		0.582**	0.226	0.182	0.293*	0.482**	0.094	0.509**	0.310*	−0.046	−0.165	−0.098	−0.039	−0.032	−0.032	0.317*	0.213	−0.004
	p-value		**0.000**	0.130	0.225	0.048	**0.001**	0.533	**0.000**	0.036	0.760	0.272	0.518	0.796	0.832	0.830	0.032	0.154	0.980
STAI_2	rho			0.309*	0.376*	0.461**	0.456**	0.085	0.678**	0.240	0.199	−0.400**	−0.025	0.202	0.062	0.310*	0.192	0.392**	−0.115
	p-value			0.036	0.010	**0.001**	**0.001**	0.576	**0.000**	0.108	0.186	0.006	0.868	0.179	0.681	0.036	0.200	0.007	0.449
ASI	rho				0.433**	0.153	0.401**	0.076	0.474**	0.377**	−0.029	−0.129	0.053	0.355*	0.576**	0.181	0.561**	0.482**	−0.008
	p-value				**0.003**	0.310	0.006	0.615	**0.001**	0.010	0.848	0.394	0.725	0.015	**0.000**	0.229	**0.000**	**0.001**	0.958
BAI	Rho					0.242	0.269	0.157	0.446**	0.505**	0.017	−0.235	0.016	0.202	0.438**	0.401**	0.368*	0.376**	0.103
	p-value					0.104	0.070	0.297	**0.002**	**0.000**	0.912	0.115	0.915	0.177	**0.002**	0.006	0.012	0.010	0.496
BDI	rho						0.362*	−0.161	0.449**	0.112	0.129	−0.254	−0.201	−0.130	0.101	0.227	0.211	0.177	−0.118
	p-value						0.013	0.286	**0.002**	0.460	0.392	0.088	0.180	0.389	0.506	0.130	0.159	0.238	0.434
PANAS	Rho							0.668**	0.668**	0.303*	0.082	−0.460**	−0.285	−0.050	0.001	0.244	0.249	0.461**	−0.029
	p-value							**0.000**	**0.000**	0.041	0.588	**0.001**	0.055	0.740	0.994	0.103	0.096	**0.001**	0.849
NEO	Rho								−0.132	0.449**	−0.348*	0.475**	0.594**	0.602**	0.289	−0.058	−0.001	0.056	0.044
	p-value								0.381	**0.002**	0.018	**0.001**	**0.000**	**0.000**	0.052	0.700	0.994	0.711	0.774
MASQ	Rho									0.111	0.335*	−0.409**	−0.174	0.130	0.111	0.381**	0.276	0.381**	−0.155
	p-value									0.463	0.023	**0.005**	0.249	0.390	0.463	0.009	0.063	0.009	0.303
TPQ	Rho										−0.367*	0.020	0.236	0.327*	0.427**	0.147	0.312*	0.429**	0.119
	p-value										0.012	0.894	0.115	0.026	**0.003**	0.328	0.035	**0.003**	0.431
BISBAS	Rho											−0.366*	−0.122	−0.152	−0.294*	0.055	−0.069	−0.140	−0.472**
	p-value											0.012	0.420	0.312	0.047	0.714	0.649	0.355	**0.001**
LotR	Rho												0.521**	0.132	0.104	−0.195	0.040	−0.101	−0.034
	p-value												**0.000**	0.383	0.492	0.195	0.789	0.503	0.824
IMI	Rho													0.428**	0.344*	0.016	0.127	−0.008	−0.128
	p-value													**0.003**	0.019	0.918	0.401	0.960	0.395
IRI	Rho														0.494**	0.179	0.133	0.201	−0.180
	p-value														**0.000**	0.233	0.378	0.181	0.231
MISS	Rho															0.357*	0.450**	0.219	0.089
	p-value															0.015	**0.002**	0.143	0.554
FOP	Rho																0.113	0.399**	−0.007
	p-value																0.456	0.006	0.966
PCS_Tot	Rho																	0.305*	0.125
	p-value																	0.039	0.407
PSWQ	Rho																		0.071
	p-value																		0.639

*STAI 1-2, State and Trait Anxiety Inventory; ASI, Anxiety Sensitivity Index; BAI, Beck Anxiety Inventory; BDI, Beck Depression Inventory; PANAS, Positive and Negative Affective Schedule; NEO, Neuroticism-Extraversion-Openness Inventory; MASQ, Mood and Anxiety Symptom Questionnaire; TPQ, Tridimensional Personality Questionnaire; BIS/BAS, Behavioral Inhibition and Behavioral Activation Scale; LotR, Life-Orientation Test-Revisited; IMI, Intrinsic Motivation Inventory; IRI, Interpersonal Reactivity Index; MISS, Multidimensional Iowa Suggestibility Scale; FOP, Fear of Pain; PCS, Pain Catastrophizing Scale; PSWQ, Penn State Worry Questionnaire; SS, Sensation Seeking. Significant results are indicated in bold. *$p < 0.05$; **$p < 0.010$.*

level of expectation related to pain changes (e.g., reductions and increases). Future large scale studies deserve to be performed in pain patient and healthy populations to better understand the connection among psychological factors, expectancies, placebo and nocebo effects.

The inclusion of an extensive battery of questionnaires related to personality factors allowed us to reveal that expectations may predict placebo and nocebo effects independently of personality factors making it a helpful tool for health care providers.

Several studies have emphasized the need for exploring the impact of personality factors as at least one of the possible ways to interpret and understand the large variability in placebo analgesic and nocebo hyperalgesic responses. To our knowledge, this is the first study that explores how distinct psychological factors can predict placebo hypolagesic responses and nocebo hyperalgesic responses, and the potential influence of personality factors in shaping positive and negative expectancies. Collectively, the complexity and variability in placebo- and nocebo-induced pain responses highlight a need to better understand the multidimensionality of pain and its modulation related to individual expectations and psychological factors. This approach provides advantages in interpreting how pain is felt and experienced.

AUTHOR CONTRIBUTIONS

LC and NC conceived the study design. NC performed the experiments. NC and LC analyzed the data and drafted the manuscript. LC revised and finalized the manuscript.

FUNDING

This research was partially supported by the University of Maryland, Baltimore (LC), the National Institute of Dental and Craniofacial Research (1R01DE025946-0; LC) and the Cooperint Internalization Program from University of Verona, Italy (NC).

ACKNOWLEDGMENTS

The authors thank Taylor Ludman, Andrea Howe, Maxie Blasini and Cynthia Renn for help with the set-up, screening of study participants or comments on the manuscript.

REFERENCES

Aslaksen, P. M., and Lyby, P. S. (2015). Fear of pain potentiates nocebo hyperalgesia. *J. Pain Res.* 8, 703–710. doi: 10.2147/JPR.S91923

Beck, A. T., Epstein, N., Brown, G., and Steer, R. A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *J. Consult. Clin. Psychol.* 56, 893–897. doi: 10.1037/0022-006x.56.6.893

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry* 4, 561–571. doi: 10.1001/archpsyc.1961.01710120031004

Carver, C. S., and White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *J. Pers. Soc. Psychol.* 67, 319–333. doi: 10.1037/0022-3514.67.2.319

Cloninger, C. R., Przybeck, T. R., and Svrakic, D. M. (1991). The tridimensional personality questionnaire: U.S. normative data. *Psychol. Rep.* 69, 1047–1057. doi: 10.2466/pr0.1991.69.3.1047

Colagiuri, B., Schenk, L. A., Kessler, M. D., Dorsey, S. G., and Colloca, L. (2015). The placebo effect: from concepts to genes. *Neuroscience* 307, 171–190. doi: 10.1016/j.neuroscience.2015.08.017

Colloca, L., and Benedetti, F. (2006). How prior experience shapes placebo analgesia. *Pain* 124, 126–133. doi: 10.1016/j.pain.2006.04.005

Colloca, L., and Benedetti, F. (2009). Placebo analgesia induced by social observational learning. *Pain* 144, 28–34. doi: 10.1016/j.pain.2009.01.033

Colloca, L., and Grillon, C. (2014). Understanding placebo and nocebo responses for pain management. *Curr. Pain Headache Rep.* 18:419. doi: 10.1007/s11916-014-0419-2

Colloca, L., Petrovic, P., Wager, T. D., Ingvar, M., and Benedetti, F. (2010). How the number of learning trials affects placebo and nocebo responses. *Pain* 151, 430–439. doi: 10.1016/j.pain.2010.08.007

Colloca, L., Sigaudo, M., and Benedetti, F. (2008). The role of learning in nocebo and placebo effects. *Pain* 136, 211–218. doi: 10.1016/j.pain.2008.02.006

Corsi, N., Emadi Andani, M., Tinazzi, M., and Fiorio, M. (2016). Changes in perception of treatment efficacy are associated to the magnitude of the nocebo effect and to personality traits. *Sci. Rep.* 6:30671. doi: 10.1038/srep30671

Costa, P. T., and McCrae, R. R. (1985). *The NEO Personality Inventory Manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Crawford, J. R., and Henry, J. D. (2004). The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* 43, 245–265. doi: 10.1348/0144665031752934

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog Select. Doc. Psychol.* 10:85.

De Pascalis, V., Chiaradia, C., and Carotenuto, E. (2002). The contribution of suggestibility and expectation to placebo analgesia phenomenon in an experimental setting. *Pain* 96, 393–402. doi: 10.1016/S0304-3959(01)00485-7

Flaten, M. A., Aslaksen, P. M., Finset, A., Simonsen, T., and Johansen, O. (2006). Cognitive and emotional factors in placebo analgesia. *J. Psychosom. Res.* 61, 81–89. doi: 10.1016/j.jpsychores.2005.12.004

Flor, H., and Turk, D. C. (1988). Chronic back pain and rheumatoid arthritis: predicting pain and disability from cognitive variables. *J. Behav. Med.* 11, 251–265. doi: 10.1007/BF00844431

Geers, A. L., Helfer, S. G., Kosbab, K., Weiland, P. E., and Landry, S. J. (2005). Reconsidering the role of personality in placebo effects: dispositional optimism, situational expectations, and the placebo response. *J. Psychosom. Res.* 58, 121–127. doi: 10.1016/j.jpsychores.2004.08.011

Geers, A. L., Helfer, S. G., Weiland, P. E., and Kosbab, K. (2006). Expectations and placebo response: a laboratory investigation into the role of somatic focus. *J. Behav. Med.* 29, 171–178. doi: 10.1007/s10865-005-9040-5

Geers, A. L., Kosbab, K., Helfer, S. G., Weiland, P. E., and Wellman, J. A. (2007). Further evidence for individual differences in placebo responding: an interactionist perspective. *J. Psychosom. Res.* 62, 563–570. doi: 10.1016/j.jpsychores.2006.12.005

Geers, A. L., Wellman, J. A., Fowler, S. L., Helfer, S. G., and France, C. R. (2010). Dispositional optimism predicts placebo analgesia. *J. Pain* 11, 1165–1171. doi: 10.1016/j.jpain.2010.02.014

Gelfland, D. M., Gelfland, S., and Rardin, M. W. (1965). Some personality factors associated with placebo responsivity. *Psychol. Rep.* 17, 555–562. doi: 10.2466/pr0.1965.17.2.555

Goubert, L., Crombez, G., and Van Damme, S. (2004). The role of neuroticism, pain catastrophizing and pain-related fear in vigilance to pain: a structural equations approach. *Pain* 107, 234–241. doi: 10.1016/j.pain.2003.11.005

Haigh, E. A., Moore, M. T., Kashdan, T. B., and Fresco, D. M. (2011). Examination of the factor structure and concurrent validity of the Langer Mindfulness/Mindlessness Scale. *Assessment* 18, 11–26. doi: 10.1177/1073191110386342

Hayes, A. F., and Preacher, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behav. Res.* 45, 627–660. doi: 10.1080/00273171.2010.498290

Huber, A., Lui, F., and Porro, C. A. (2013). Hypnotic susceptibility modulates brain activity related to experimental placebo analgesia. *Pain* 154, 1509–1518. doi: 10.1016/j.pain.2013.03.031

Hunter, T., Siess, F., and Colloca, L. (2014). Socially induced placebo analgesia: a comparison of a pre-recorded versus live face-to-face observation. *Eur. J. Pain* 18, 914–922. doi: 10.1002/j.1532-2149.2013.00436.x

Johnston, N. E., Atlas, L. Y., and Wager, T. D. (2012). Opposing effects of expectancy and somatic focus on pain. *PLoS ONE* 7:e38854. doi: 10.1371/journal.pone.0038854

Kotov, R. I., Bellman, S. B., and Watson, D. B. (2004). *Multidimensional Iowa Suggestibility Scale (MISS)* (New York, NY: Stony Brook University).

Kudel, I., Edwards, R. R., and Moric, M. (2005). The role of neuroticism, pain catastrophizing and pain-related fear in vigilance to pain: a structural equations approach. A comment on Goubert et al. (2004). *Pain* 115, 214–216; author reply 16–17. doi: 10.1016/j.pain.2005.02.008

Kube, T., Rief, W., and Glombiewski, J. A. (2017). On the Maintenance of Expectations in Major Depression - Investigating a Neglected Phenomenon. *Front. Psychol.* 8:9. doi: 10.3389/fpsyg.2017.00009

Lui, F., Colloca, L., Duzzi, D., Anchisi, D., Benedetti, F., and Porro, C. A. (2010). Neural bases of conditioned placebo analgesia. *Pain* 151, 816–824. doi: 10.1016/j.pain.2010.09.021

Lund, K., Petersen, G. L., Erlandsen, M., De Pascalis, V., Vase, L., Jensen, T. S., et al. (2015). The magnitude of placebo analgesia effects depends on how they are conceptualized. *J. Psychosom. Res.* 79, 663–668. doi: 10.1016/j.jpsychores.2015.05.002

Lyby, P. S., Aslaksen, P. M., and Flaten, M. A. (2010). Is fear of pain related to placebo analgesia? *J. Psychosom. Res.* 68, 369–377. doi: 10.1016/j.jpsychores.2009.10.009

Markland, D., and Hardy, L. (1997). On the factorial and construct validity of the Intrinsic Motivation Inventory: conceptual and operational concerns. *Res. Q. Exerc. Sport* 68, 20–32. doi: 10.1080/02701367.1997.10608863

Mehta, S., Rice, D., Janzen, S., Pope, J. E., Harth, M., Shapiro, A. P., et al. (2016). Mood, disability, and quality of life among a subgroup of rheumatoid arthritis individuals with experiential avoidance and anxiety sensitivity. *Pain Res. Manag.* 2016:7241856. doi: 10.1155/2016/7241856

Meyer, T. J., Miller, M. L., Metzger, R. L., and Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behav. Res. Ther.* 28, 487–495. doi: 10.1016/0005-7967(90)90135-6

Morton, D. L., Watson, A., El-Deredy, W., and Jones, A. K. (2009). Reproducibility of placebo analgesia: effect of dispositional optimism. *Pain* 146, 194–198. doi: 10.1016/j.pain.2009.07.026

Nes, L. S., and Segerstrom, S. C. (2006). Dispositional optimism and coping: a meta-analytic review. *Pers. Soc. Psychol. Rev.* 10, 235–251. doi: 10.1207/s15327957pspr1003_3

Ober, K., Benson, S., Vogelsang, M., Bylica, A., Günther, D., Witzke, O., et al. (2012). Plasma noradrenaline and state anxiety levels predict placebo response in learned immunosuppression. *Clin. Pharmacol. Ther.* 91, 220–226. doi: 10.1038/clpt.2011.214

Osman, A., Breitenstein, J. L., Barrios, F. X., Gutierrez, P. M., and Kopper, B. A. (2002). The Fear of Pain Questionnaire-III: further reliability and validity with nonclinical samples. *J. Behav. Med.* 25, 155–173. doi: 10.1023/A:1014884704974

Peciña, M., Azhar, H., Love, T. M., Lu, T., Fredrickson, B. L., Stohler, C. S., et al. (2013). Personality trait predictors of placebo analgesia and neurobiological correlates. *Neuropsychopharmacology* 38, 639–646. doi: 10.1038/npp.2012.227

Petrovic, P. (2008). Placebo analgesia and nocebo hyperalgesia–two sides of the same coin? *Pain* 136, 5–6. doi: 10.1016/j.pain.2008.03.004

Preacher, K. J., and Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav. Res. Methods Instrum. Comput.* 36, 717–731. doi: 10.3758/BF03206553

Reiss, S., Peterson, R. A., Gurskey, D. M., and McNally, R. J. (1986). Anxiety sensitivity, anxiety frequency, and the prediction of fearfulness. *Behav. Res. Ther.* 24, 1–8. doi: 10.1016/0005-7967(86)90143-9

Rütgen, M., Seidel, E. M., Riecanský, I., and Lamm, C. (2015a). Reduction of empathy for pain by placebo analgesia suggests functional equivalence of empathy and first-hand emotion experience. *J. Neurosci.* 35, 8938–8947. doi: 10.1523/JNEUROSCI.3936-14.2015

Rütgen, M., Seidel, E. M., Silani, G., Riecanský, I., Hummer, A., Windischberger, C., et al. (2015b). Placebo analgesia and its opioidergic regulation suggest that empathy for pain is grounded in self pain. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5638–E5646. doi: 10.1073/pnas.1511269112

Scheier, M. F., Carver, C. S., and Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. *J. Pers. Soc. Psychol.* 67, 1063–1078. doi: 10.1037/0022-3514.67.6.1063

Schweinhardt, P., Seminowicz, D. A., Jaeger, E., Duncan, G. H., and Bushnell, M. C. (2009). The anatomy of the mesolimbic reward system: a link between personality and the placebo analgesic response. *J. Neurosci.* 29, 4882–4887. doi: 10.1523/JNEUROSCI.5634-08.2009

Severeijns, R., Vlaeyen, J. W., van den Hout, M. A., and Weber, W. E. (2001). Pain catastrophizing predicts pain intensity, disability, and psychological distress independent of the level of physical impairment. *Clin. J. Pain* 17, 165–172. doi: 10.1097/00002508-200106000-00009

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory STAI (form Y)*. Palo Alto, CA: Consulting Psychologist Press Inc.

Staats, P. S., Staats, A., and Hekmat, H. (2001). The additive impact of anxiety and a placebo on pain. *Pain Med.* 2, 267–279. doi: 10.1046/j.1526-4637.2001.01046.x

Sullivan, M. J., Bishop, S. R., and Pivik, J. (1995). The pain catastrophizing scale: development and validation. *Psychol. Assess.* 7, 524–532. doi: 10.1037/1040-3590.7.4.524

Swider, K., and Babel, P. (2013). The effect of the sex of a model on nocebo hyperalgesia induced by social observational learning. *Pain* 154, 1312–1317. doi: 10.1016/j.pain.2013.04.001

Trost, Z., Strachan, E., Sullivan, M., Vervoort, T., Avery, A. R., and Afari, N. (2015). Heritability of pain catastrophizing and associations with experimental pain outcomes: a twin study. *Pain* 156, 514–520. doi: 10.1097/01.j.pain.0000460326.02891.fc

Vogtle, E., Barke, A., and Kroner-Herwig, B. (2013). Nocebo hyperalgesia induced by social observational learning. *Pain* 154, 1427–1433. doi: 10.1016/j.pain.2013.04.041

Zubieta, J. K., Yau, W. Y., Scott, D. J., and Stohler, C. S. (2006). Belief or Need? Accounting for individual variations in the neurochemistry of the placebo effect. *Brain Behav. Immun.* 20, 15–26. doi: 10.1016/j.bbi.2005.08.006

Zuckerman, M. (1994). *Behavioral Expressions on Biosocial Bases of Sensation-Seeking*. New York, NY: Cambridge University Press.

Conflict of Interest Statement: LC has received lecture honoraria (Georgetown University and Stanford University) and has acted as speaker or consultant for Grünenthal and Emmi Solution. NC has no conflicts of interest to be declared.

The reviewer JG and the handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2017 Corsi and Colloca. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

REVIEW
published: 21 February 2017
doi: 10.3389/fpsyg.2017.00233

Patients' Expectations Regarding Medical Treatment: A Critical Review of Concepts and Their Assessment

Johannes A. C. Laferton[1,2], Tobias Kube[2], Stefan Salzmann[2], Charlotte J. Auer[3] and Meike C. Shedden-Mora[4]*

[1] Department of Psychology, Clinical Psychology and Psychotherapy, Psychologische Hochschule Berlin, Berlin, Germany, [2] Department of Psychology, Division of Clinical Psychology and Psychotherapy, Philipps University of Marburg, Marburg, Germany, [3] Division of Psychotherapy and Psychiatry, University Hospital Lübeck, Lübeck, Germany, [4] Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Paul Van Schaik, Teesside University, UK
Annelie Rosén, Karolinska Institutet, Sweden

***Correspondence:**
Johannes A. C. Laferton
j.laferton@psychologische-hochschule.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *15 November 2016*
Accepted: *06 February 2017*
Published: *21 February 2017*

Citation:
Laferton JAC, Kube T, Salzmann S, Auer CJ and Shedden-Mora MC (2017) Patients' Expectations Regarding Medical Treatment: A Critical Review of Concepts and Their Assessment. Front. Psychol. 8:233. doi: 10.3389/fpsyg.2017.00233

Patients' expectations in the context of medical treatment represent a growing area of research, with accumulating evidence suggesting their influence on health outcomes across a variety of medical conditions. However, the aggregation of evidence is complicated due to an inconsistent and disintegrated application of expectation constructs and the heterogeneity of assessment strategies. Therefore, based on current expectation concepts, this critical review provides an integrated model of patients' expectations in medical treatment. Moreover, we review existing assessment tools in the context of the integrative model of expectations and provide recommendations for improving future assessment. The integrative model includes expectations regarding treatment and patients' treatment-related behavior. Treatment and behavior outcome expectations can relate to aspects regarding benefits and side effects and can refer to internal (e.g., symptoms) and external outcomes (e.g., reactions of others). Furthermore, timeline, structural and process expectations are important aspects with respect to medical treatment. Additionally, generalized expectations such as generalized self-efficacy or optimism have to be considered. Several instruments assessing different aspects of expectations in medical treatment can be found in the literature. However, many were developed without conceptual standardization and psychometric evaluation. Moreover, they merely assess single aspects of expectations, thus impeding the integration of evidence regarding the differential aspects of expectations. As many instruments assess treatment-specific expectations, they are not comparable between different conditions. To generate a more comprehensive understanding of expectation effects in medical treatments, we recommend that future research should apply standardized, psychometrically evaluated measures, assessing multidimensional aspects of patients' expectations that are applicable across various medical treatments. In the future, more research is needed on the interrelation of different expectation concepts as well as on factors influencing patients' expectations of illness and

treatment. Considering the importance of patients' expectations for health outcomes across many medical conditions, an integrated understanding and assessment of such expectations might facilitate interventions aiming to optimize patients' expectations in order to improve health outcomes.

Keywords: expectations, outcome expectancy, self-efficacy, optimism, placebo effect, treatment, assessment, operationalization

INTRODUCTION

The relevance of patients' expectations for health outcomes has received increasing attention in recent years. Expectations play an important role in both physical (Di Blasi et al., 2001; Mondloch et al., 2001) and mental health (Constantino et al., 2011; Rief et al., 2015; Kube et al., 2017). Moreover, they are a key mechanism of the placebo and nocebo effect, a phenomenon according to which subjective and physiological changes emerge due to inert or non-specific treatment components (Colloca and Miller, 2011b; Enck et al., 2013). Accumulating evidence suggests that expectations influence treatment outcome in patients with various medical conditions. For instance, they have been linked to course and treatment outcome in patients with heart disease (Petrie et al., 1996; Juergens et al., 2010; Barefoot et al., 2011; Habibovic et al., 2014), stroke (Jones and Riazi, 2011), cancer (Colagiuri and Zachariae, 2010; Nestoriuc et al., 2016), musculoskeletal disorders (Mahomed et al., 2002; Oettingen and Mayer, 2002; van den Akker-Scheek et al., 2007), injuries (Booth-Kewley et al., 2014; Murgatroyd et al., 2016) and obesity (Oettingen and Wadden, 1991; Armitage et al., 2015; Crane et al., 2016). Expectations even predict outcome in patients undergoing different kinds of surgery (Auer et al., 2016a). Hence, patients with more positive expectations seem to be more likely to benefit from medical treatment across medical conditions.

However, despite the growing number of studies investigating expectations in different medical conditions, it is difficult to integrate current findings. The heterogeneity with regard to the conceptualization and assessment of patients' expectations (van Hartingsveld et al., 2010; Bowling et al., 2012; Zywiel et al., 2013) has been considered as a major limitation in several systematic reviews and meta-analyses (Mondloch et al., 2001; Fadyl and McPherson, 2008; Haanstra et al., 2012; Auer et al., 2016a). Some theoretical concepts refer to overlapping aspects of expectations using different terminology, which further complicates the integration of evidence regarding patients' expectations (Maddux, 2007). Moreover, many studies focus on a single or only a few aspects of expectations, making it difficult to investigate the differential influence of distinct expectation concepts (Haanstra et al., 2015b; Laferton et al., 2015a; Auer et al., 2016b).

Unambiguous terminology, conceptual integration, and standardized assessment are required in order to foster understanding and clinically harness the relationship between expectations and health. The current review has two aims. First, based on a review of current expectation concepts, we aim to provide an integrated model of patients' expectations in medical treatment. Second, we review the most relevant existing assessment tools and provide recommendations for improving the assessment of expectations with the aim of facilitating more integrative and standardized future research.

PATIENTS' EXPECTATIONS REGARDING MEDICAL TREATMENT: AN OVERVIEW OF CONCEPTS

Expectations are among the most studied constructs in psychological research and have been explicitly or implicitly embedded in many psychological theories (Maddux, 1999). There are many types of expectations in the literature with often ambiguous terminology (Bowling et al., 2012). In the following, theoretical concepts and aspects of patients' expectations, which are of relevance for health outcomes in medical treatment contexts, are reviewed. They are summarized within an integrative model of expectations of patients undergoing medical treatment (see **Figure 1**) to facilitate an unambiguous and more integrated use of terminology and concepts.

In this manuscript, the term patients' expectations refers to future-directed beliefs that focus on the incidence or non-incidence of a specific event or experience (Kube et al., 2016). They can manifest as conscious future-directed cognitions, or they may be present without full awareness (e.g., in the case of conditioned learning processes; Kirsch, 2004; Kirsch et al., 2004, 2014). In this sense, expectations are of a predictive nature and need to be distinguished from constructs that have been termed *ideal expectations*, *value expectations* or *fantasies* (Kravitz, 1996; Oettingen and Mayer, 2002; David et al., 2004; Leung et al., 2009). The latter constructs refer to what a patient would like to happen and are more an expression of hopes or desires than a probabilistic estimation about the future. Ideal expectations or fantasies seem to have opposite effects on health outcomes when compared with patients' predictive expectations, which empirically confirms the differentiation between the two constructs (Oettingen and Wadden, 1991; Oettingen and Mayer, 2002; Kappes and Oettingen, 2011; Johannessen et al., 2012; Oettingen, 2012).

The following overview of expectation concepts includes social learning and social cognitive theories, the response expectancy theory, the common sense model of illness representation, as well as a short summary of other expectation dimensions. Importantly, our review does not claim to be exhaustive, but rather aims to integrate the most relevant theoretical concepts.

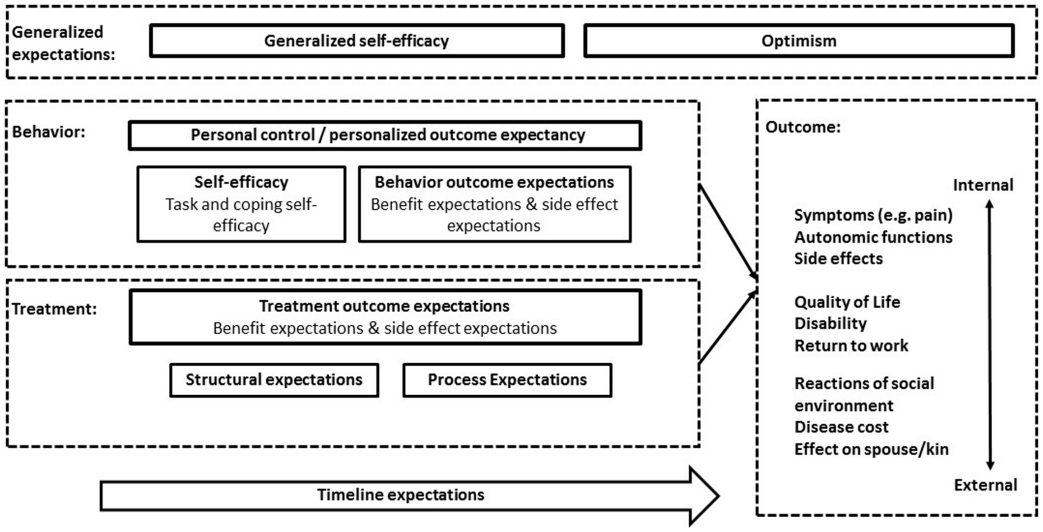


FIGURE 1 | Schematic illustration of the integrative model of expectations in patients undergoing medical treatment. Behavior, treatment and outcome related aspects of expectations refer to the specific treatment context. Generalized expectations are independent of the specific treatment context, but might influence specific expectations and treatment outcome. Timeline expectations refer to temporal aspects of the disease, treatment and health behavior, e.g., the course of disease in the context of the treatment.

Social Learning and Social Cognitive Theories

Among the most prominent theoretical backgrounds for the conceptualization of expectations are social learning and social cognitive theories (Bandura, 1986; Maddux, 1999; Schwarzer, 1994), which distinguish two main concepts of expectations: (1) *Behavior outcome expectancies* express the (subjective) likelihood that a specific outcome will follow a given action (e.g., regular exercise will lead to health benefits). These outcomes can be of a physical, social or self-evaluative nature (Bandura, 1997); (2) *self-efficacy* expresses an individual's expectation of being capable of executing a certain action (e.g., ability to exercise regularly). Self-efficacy can be further distinguished into *task self-efficacy* and *coping self-efficacy* (Kirsch, 1995). While the former expresses the perceived ability to perform a particular behavior (e.g., being able to perform a specific exercise, e.g., jogging), the latter refers to the ability to prevent, control or cope with the demands that might be experienced when performing the behavior (e.g., being able to motivate oneself for regular exercise or being able to tolerate exercise-induced exhaustion). Self-efficacy and behavior outcome expectations play an important role in volitional agentic behavior (Bandura, 2001). However, they do not fully account for the relationship between expectations and non-volitional responses to treatment (Maddux, 1999), such as cardiovascular functions, immune and endocrine functions or pain, as shown by research on the placebo effect (Price et al., 2008; Enck et al., 2013). Non-volitional responses are especially important for expectations regarding medical treatments. Although patient behavior such as medication adherence (Sokol et al., 2005) or a healthy lifestyle (Willett, 2002) plays an important role in medical conditions, in most medical treatments, the patient is largely a responder to external stimuli (e.g., medication, surgical procedures, manual therapy, radiation).

Response Expectancy Theory

Kirsch's (1983, 1997) response expectancy theory adds further important aspects of expectations, differentiating between stimulus expectancies and response expectancies. Accordingly, with regard to the outcome that is expected to occur, Kirsch distinguishes between expected external/environmental outcomes (*stimulus expectancies)* and expected non-volitional, internal outcomes (*response expectancies)*. He argues that most theories of expectations are concerned with stimulus expectancies, such as the expectation of money or recognition by others as a result of a certain behavior (Kirsch, 1983). Response expectancies, on the other hand, refer to the expected occurrence of the individual's non-volitional, internal responses to a certain external stimulus (e.g., the expectation that an analgesic will lead to pain reduction) or to one's own behavior (e.g., the expectation that a relaxation exercise will reduce subjective stress). Thus, response expectancies cover both aspects of medical treatment: the patient as a passive recipient of medical treatment and the patient's volitional health-directed behavior. Moreover, expectations regarding non-volitional responses such as change in symptoms or autonomic bodily functions are of outmost importance for patients with medical conditions, as they are often the focus of the disease experience.

Common Sense Model of Illness Representation

According to the common sense model of illness representation (Leventhal et al., 1980), patients have subjective models about their illness, which comprise interrelated beliefs about the illness

and its effect on their lives (Petrie and Weinman, 2012). These beliefs are related to important health outcomes in a broad range of medical conditions (Hagger and Orbell, 2003; Petrie et al., 2007). A patient's illness perceptions include beliefs about what caused the illness (*causes*), how long it will last (*timeline*), the *consequences* for the patient's life, which symptoms are attributed to the illness (*identity*), and how the condition can be controlled or cured by the patient's behavior (*personal control*) or by the treatment (*treatment control*). Although the common sense model does not include expectations as an explicitly denoted construct, expectations are conceptualized as a major underlying component of the different beliefs (Cameron and Leventhal, 2003). For instance, expectations are an inherent part of illness beliefs, including the prediction of future events or experiences, thus referring to timeline, personal control and treatment control as well as (future) consequences. In this regard, the common sense model covers important dimensions of patients' expectations related to their illness and treatment.

Additional Dimensions of Expectations

Several other aspects of expectations have been mentioned in the literature (Bowling et al., 2012). *Process* or *structural expectations* (e.g., sequence of steps in a treatment procedure; shape and color of a medication; a physician's treatment ritual) are an important part of the context in which a treatment takes place, which in turn is a major factor in the placebo effect (Di Blasi et al., 2001; Colloca and Miller, 2011a). Expectations about the structural or process-related aspects of a treatment are likely to influence outcome expectations. For example, expectation effects for the same analgesic are higher when it is applied via a syringe rather than in pill form (de Craen et al., 2000) or when it is openly administered by a physician compared to hidden administration by an automatic device (Price et al., 2008). Similarly, cardiac patients have higher outcome expectations for more invasive procedures (Hirani et al., 2008).

A more self-evident aspect is the *valence* of patients' expectations. This can be conceptualized either on one dimension, namely expectations of high vs. low *treatment benefit* (e.g., expectation that a treatment will relieve all pain vs. some pain), or on two relatively independent dimensions, namely expectations of treatment benefit and treatment-related *side effects* (e.g., expecting that a treatment will lead to both pain relief and distressing side effects like nausea). Negative expectations about side effects or adverse events can themselves induce the experience of nocebo-related side effects (Barsky et al., 2002; Colloca and Finniss, 2012). Moreover, distinct positive and negative dimensions also apply to behavior outcome expectations (Schwarzer, 1994), e.g., conceptualized as cost and benefit expectations in the *Health Belief Model* (Becker, 1974).

Expectations can further vary in their degree of *specificity* or *generalization*, meaning that they can be held for very specific contexts only (e.g., a specific treatment for a specific medical condition), for several similar contexts (e.g., a specific medical condition or a specific treatment), or ultimately any situation. The most prominent generalized outcome expectation is the concept of *dispositional optimism* (Carver et al., 2010; Hanssen et al., 2013), which has been extensively linked to favorable health outcomes. Notably, dispositional optimism has also been associated with an enhanced placebo response (Geers et al., 2010). In a similar vein, self-efficacy expectations can be context-specific, domain-specific or can ultimately be applied to a broad range of behaviors, as conceptualized in the concept of *generalized self-efficacy* (Schwarzer, 1994; Schwarzer and Jerusalem, 1995).

Other aspects include the *strength of expectations* and their *relation to reality*. The former refers to how strongly a person is convinced of his/her expectation, hence resembling a subjective reality. The latter is a judgment about how realistic an expectation actually is or was. This can only be assessed *post hoc*, or might be estimated based on existing empirical findings or expert judgments.

Integrative Model of Expectations in Patients Undergoing Medical Treatment

To summarize, several aspects have to be considered for an integrative model of expectations in patients undergoing medical treatment (see **Figure 1**). Expectations can either be related to a patient's illness- and treatment-related behavior or to the treatment the patient is receiving (Crow et al., 1999; van Hartingsveld et al., 2010). However, contrary to previous conceptualizations (Crow et al., 1999), which considered self-efficacy as the only aspect of expectations regarding patient behavior, one can argue that behavior-related expectations should be divided into self-efficacy and behavior outcome expectations. A patient with high self-efficacy for engaging in regular physical exercise will not start exercising unless he/she also expects exercising to lead to health benefits (behavior outcome expectation). The combination of self-efficacy and behavior outcome expectations has been termed *personalized outcome expectancy* (Kirsch, 1995) or *personal control beliefs* (Cameron and Leventhal, 2003). Treatment-related expectations consist of expectations regarding treatment outcome as well as the structural and process-related aspects of the treatment (Haanstra et al., 2013), which are likely to influence treatment outcome expectations. Both behavioral and treatment outcome expectations can refer to distinguishable expectations of benefits and side effects. Moreover, the expected outcome of a behavior or treatment can be distinguished into the two basic categories described above: (1) expectations of non-volitional, internal changes such as symptoms or autonomic functions, and (2) external expectancies, referring to the expectations of external changes such as reactions of the social environment. Moreover, patients hold expectations about the temporal dimension of their behavior, treatment, disease and the expected outcomes (*timeline expectations*). Finally, it is necessary to consider generalized expectations, such as generalized self-efficacy and generalized outcome expectations (optimism), as these have been shown to influence outcome and are likely to influence specific aspects of expectations in patients undergoing medical treatment (Schwarzer, 1994; Carver et al., 2010).

OPERATIONALIZATION OF EXPECTATIONS

The proposed model of expectations of patients undergoing medical treatment not only aims to resolve ambiguity on a theoretical level, but also applies to the assessment and therefore the reporting of results on expectation effects. To facilitate the aggregation of evidence on differential aspects of expectations, the model seeks to foster a consistent operationalization and assessment of expectation constructs. In many studies that do not rely on precise terminology and explicit theoretical concepts, these issues can only be detected by inspecting the original items used in the expectation assessment (Kirsch, 1995). The use of the conceptual distinctions of expectations and their precise terminology reviewed in this manuscript should facilitate the resolution of such issues in future research. In the following, examples of instruments assessing expectations in patients undergoing medical treatment are classified in the context of the proposed integrative model of expectations. Subsequently, several issues of the current practice of expectation assessment are pointed out to encourage the advancement of future operationalization.

Overview of Assessment Instruments According to the Integrative Model of Expectations

Given the aforementioned heterogeneity of assessment instruments, it is beyond the scope of the present work to provide an exhaustive review of assessment instruments for expectations in the medical treatment context. More importantly, in the following paragraph, we will review instruments of relevance to the integrative model of patients' expectations. **Table 1** identifies the expectation dimensions that are assessed by the outlined instruments.

Multidimensional Instruments

The instrument that assesses the broadest range of expectation aspects using distinguishable scales is the Revised Illness Perceptions Questionnaire (IPQ-R; Moss-Morris et al., 2002) and its short form (Brief Illness Perceptions Questionnaire; B-IPQ; Broadbent et al., 2006). This very well established instrument offers the possibility to distinguish between treatment control expectations, personal control expectations, timeline expectations and, if reformulated to refer to the future, expected consequences (McCarthy et al., 2003; Laferton et al., 2013), thus satisfying the required multidimensional assessment of expectations.

Mixed-Dimensional Instruments

As shown in **Table 1**, most assessment instruments are not specific to a certain concept of the integrative model of expectations, and many of them aggregate items in relation to several dimensions within one expectation score. For instance, the Future Expectations Regarding Life with Heart Disease scale (FERLHDS; Axelrad, 1982) has been used several times in patients with heart disease and has shown acceptable internal consistency as well as construct and predictive validity (Brummett et al., 2004; Chunta, 2009; Barefoot et al., 2011). The measure has recently been adapted for patients undergoing cardiac surgery, again with acceptable reliability and validity (C-SPEQ; Holmes et al., 2016). Both scales use items assessing behavior- and treatment-related expectations with respect to disease-specific and more general expected outcome that are either positively or negatively framed and concern both internal and external outcome expectations. Furthermore, singular items refer to process and to some extent timeline expectations. All 18 items are summed up to form a single expectation score. Additionally, the Positive Health Expectations Scale (PHES; Leedham et al., 1995) has been used in several cardiac surgery populations (Leedham et al., 1995; Sears et al., 2004; Auer et al., 2016b); its internal consistency as well as construct and predictive validity have been confirmed. The scale primarily assesses treatment outcome expectations in relation to more general outcome dimensions such as general physical functioning and quality of life. Additional items ask about motivational aspects and general outlook on life. Again, all items are integrated into a single expectation score.

Unidimensional Instruments

Given the impact of social learning theories, self-efficacy has been more frequently operationalized on an explicit theoretical basis compared to most other aspects of patients' expectations (Bowling et al., 2012). Specific self-efficacy has been assessed in relation to various medical conditions and health behaviors (e.g., Holden, 1991), leading to a large number of specific self-efficacy instruments, for instance for walking (Jenkins and Gortner, 1998), physical exercise (e.g., Schwarzer et al., 2008), nutrition behaviors (Schwarzer and Renner, 2016) or rehabilitation behavior (Waldrop et al., 2001). An exhaustive review of specific self-efficacy instruments is beyond the scope of this manuscript. Only a small number of instruments incorporate both aspects of behavior-related expectations: self-efficacy and behavior outcome expectations. The parallel assessment of both constructs is not indicated if the outcome is largely determined by one's behavior (Maddux, 1999). However, if this is not the case, it might be valuable to measure personalized outcome expectations or to assess both self-efficacy and behavior outcome expectations. For example, Dougherty et al. (2007) developed a scale that assesses both self-efficacy and behavior outcome expectations in patients undergoing cardioverter defibrillator implantation. Besides the IPQ scales, another instrument assessing the aspect of perceived personal control is the Control Attitudes Scale (CAS; Moser and Dracup, 1995) and its revised form (CAS-R; Moser et al., 2009), which has been psychometrically evaluated in cardiac patients.

Furthermore, several instruments assess generalized expectation constructs. The Life Orientation Test and its revised version (LOT-R; Scheier and Carver, 1985; Scheier et al., 1994), which assess dispositional optimism, constitute a standardized measure that has been extensively evaluated and which further provides population-based norm values (Glaesmer et al., 2012). Moreover, generalized self-efficacy can be assessed with a standardized, psychometrically well-evaluated

TABLE 1 | Overview of instruments with regard to the aspects of the integrative model of expectations in patients undergoing medical treatment.

Instrument	Expectation construct	Dimensionality	Generic/specific
IPQ-R/B-IPQ (Moss-Morris et al., 2002; Broadbent et al., 2006)	• Personal control (s) • Treatment outcome (s) • Timeline (s) • Consequences (s; if formulated toward the future; see McCarthy et al., 2003; Laferton et al., 2013)	Multi	Generic
FERLHDS (Axelrad, 1982) and C-SPEQ (Holmes et al., 2016)	• Personalized outcome expectancy (i) • Treatment outcome (i) • Process (i) • Timeline (i)	Mixed	Specific
PHES (Leedham et al., 1995)	• Treatment outcome (i) • Timeline (i) • Optimism (i)	Mixed	Generic
SE-ICD and OE-ICD (Dougherty et al., 2007)	• Self-efficacy (s) • Behavior outcome expectations (s)	Multi	Specific
CAS-R (Moser et al., 2009)	• Personalized outcome expectancy (s; "perceived control")	Single	Specific
LOT-R (Scheier and Carver, 1985)	• Generalized outcome expectancy (s; optimism and pessimism)	Single	Generic
GSE (Schwarzer and Jerusalem, 1995)	• Generalized self-efficacy (s)	Single	Generic
MODEMS (Tashjian et al., 2007)	• Treatment outcome (s)	Single	Specific
NKSSS (Noble et al., 2012)	• Treatment outcome (s)	Single	Specific
PDI-E (Laferton et al., 2013)	• Treatment outcome (s)	Single	Generic
ADL-E (Dohnke et al., 2006)	• Treatment outcome (s)	Single	Generic
PCS-E (Powell et al., 2012)	• Treatment outcome (s)	Single	Generic
CEQ (Devilly and Borkovec, 2000)	• Treatment outcome (s)	Single	Generic
EXPECT-ICD (Habibovic et al., 2014)	• Positive treatment outcome (s); • Negative treatment outcome (s)	Multi	Specific
GASE-EXPECT (von Blanckenburg et al., 2013)	• Negative treatment outcome (s)	Single	Generic
ANP-E (Hüppe et al., 2013)	• Negative treatment outcome (s)	Single	Specific

(s) = aspect represented by independent scale. (i) = aspect represented by singular item. Dimensionality: Multi = Several expectation dimensions are each assessed by an independent scale; Mixed = Several expectation dimensions are assessed by single items that are subsumed in one scale; Single = Only one expectation dimension is assessed.

instrument, the Generalized Self-Efficacy Scale (GSE; Schwarzer and Jerusalem, 1995).

Regarding treatment outcome expectations, a frequent strategy is to adapt instruments or criteria which are commonly used to assess treatment outcome. Following this strategy, some instruments incorporate disease-specific treatment outcome expectations, such as the expectation module of the Musculoskeletal Outcomes Data Evaluation and Management System (MODEMS; Tashjian et al., 2007) or the expectation module of the New Knee Society Scoring System (NKSSS; Noble et al., 2012). Similarly, studies investigating placebo effects have assessed expectations in terms of expected treatment outcome (Bingel et al., 2011; Kirsch et al., 2014).

Other instruments assess treatment outcome expectations by exclusively asking about generic outcome dimensions such as disability, return to work or quality of life. The Pain Disability Index (Tait et al., 1990) has been recently adapted (PDI-E; Laferton et al., 2013) to assess expected disability in seven areas of daily living. So far, it has been used in two independent studies assessing expectations of peripheral arterial disease (Ferrari et al., 2015) or heart surgery (Rief et al., 2017). It was shown to be have good internal consistency (Laferton et al., 2015b) and construct validity (Laferton et al., 2015a). In a similar fashion, Dohnke et al. (2006) assessed expectations for activities of daily living (ADL-E) in hip joint replacement rehabilitation patients. Powell et al. (2012) assessed expectations by adapting the SF-36 physical functioning quality of life component score (PCS-E), although both of the aforementioned studies failed to report the psychometric evaluation of the scales. Another generic way to assess patients' expectations is to ask about their perceived likelihood of return to work (Fadyl and McPherson, 2008), which is highly relevant for many patients. Finally, the Credibility Expectancy Questionnaire (CEQ; Devilly and Borkovec, 2000) is an evaluated and frequently used instrument to assess patients' perceived treatment credibility and treatment outcome expectations on a generic level. Originally, the CEQ was developed for application within psychotherapeutic treatment, but it can be easily adapted for the medical treatment context (e.g., Haanstra et al., 2015a,b).

Few instruments exist for the specific assessment of negative outcome or side-effect expectations. The EXPECT-ICD (Habibovic et al., 2014) assesses positive and negative treatment outcome expectations of patients undergoing cardioverter defibrillator device implantation. The scale includes items

assessing both disease-specific outcome dimensions and more generalized outcome dimensions such as physical functioning and quality of life. Moreover, some instruments specifically assess side-effect expectations for pharmacological treatment. The General Assessment of Side Effects Scale (Rief et al., 2011) assesses the most common medication side effects and has recently been adapted for the assessment of expectations about side effects of breast cancer patients undergoing endocrine therapy (GASE-EXPECT; von Blanckenburg et al., 2013). It has shown good initial internal consistency and validity (Heisig et al., 2015; Nestoriuc et al., 2016) and can be adapted to incorporate medication-specific symptoms. In a similar vein, Hüppe et al. (2013) assessed expectations for general anesthesia-related side effects by adapting the Anaesthesiological Questionnaire (ANP-E; Hüppe et al., 2003) for the measurement of side effects. Moreover, several measurement instruments have been developed based on the common sense model of illness representation. These instruments incorporate treatment concerns, which combine expectations about side effects with more general aspects of worrying in the context of treatment. The subscale "concerns" of the Beliefs about Medicines Questionnaire (Horne et al., 1999) incorporates expectations about negative effects of medications. Similar instruments have also been developed to assess concerns about surgery (Francis et al., 2009) or heart disease treatment (Hirani et al., 2008).

In sum, although some standardized measurements have been developed to assess different aspects of expectations, very few studies have examined the extent to which these different measures conceptually overlap (e.g., Haanstra et al., 2015b; Laferton et al., 2015a; Auer et al., 2016b; Heisig et al., 2016). Despite this variety of assessment instruments, the current practice of assessing patients' expectations in the medical treatment context can be further improved. In the following, we provide recommendations for improving the future assessment of expectations in patients undergoing medical treatment.

Recommendations for Improving the Assessment of Expectations in Patients Undergoing Medical Treatment

Standardized Assessment

Several reviews concluded that there is a lack of standardized assessment of medical patients' expectations (Fadyl and McPherson, 2008; Haanstra et al., 2012; Auer et al., 2016a). Besides lacking conceptual standardization as discussed above, many instruments were developed and used for only one investigation, often without providing a rationale for development or data on psychometric evaluation (van Hartingsveld et al., 2010; Bowling et al., 2012; Zywiel et al., 2013). This is a major issue, as without knowledge about reliability and validity, the evidence collected using such an instrument is subject to major limitations. To gather more credible evidence, measurement instruments need to be developed based on a transparent rationale. Possible strategies may include theory-guided development, qualitative research on patients' expectations, expert focus groups or the adaptation of well-developed patient-reported outcome tools. Further, the dimensionality of the measurement tool not only needs to be developed in an exploratory manner, but also needs to be tested in a confirmatory manner in independent samples. Moreover, reliability, construct validity and predictive validity need to be confirmed across several studies.

Multidimensional Assessment

A further issue is the lack of multidimensionality. Many studies merely assess one aspect of expectations (e.g., behavior- vs. treatment-related expectations; van Hartingsveld et al., 2010; Zywiel et al., 2013). If one wishes to assess the expectation effects in relation to a single application of an analgesic (e.g., in an experimental investigation of placebo effects), the assessment of treatment-related expectations might cover most of the relevant expectations in that context. The same might apply to studies investigating expectation effects related to patient behavior in the absence of any treatment. However, this hinders the collection of integrative evidence regarding the predictive value of distinct aspects of expectations in medical conditions (see also Auer et al., 2016a,b). This is also problematic for clinical practice, as for the majority of patients with medical conditions, several aspects of expectations appear to be important (e.g., expectations about treatment efficacy, personal control over as well as consequences of a particular disease; Haanstra et al., 2013). Measuring only one aspect does not cover the whole picture. Similarly, if several aspects of expectations were assessed at the same time, but were not distinguished by separate (sub-)scales of the instrument, this would impede knowledge about the differential role of certain aspects of expectations. Therefore, the parallel application of instruments measuring different aspects of expectations or the use of an instrument distinguishing certain aspects of expectations is essential. The parallel assessment of the dimensions listed in the following paragraphs should be especially considered when assessing medical patients' expectations.

As mentioned above, in most medical treatment contexts, both the patients' illness- and treatment-related behavior and the treatment itself are important factors for treatment success (Crow et al., 1999; van Hartingsveld et al., 2010). Therefore, both treatment- and behavior-related expectations are likely to influence health outcomes. Yet, very few instruments incorporate separate scales for both aspects of expectations (see **Table 1**) and only a small number of studies use separate instruments to measure both treatment- and behavior-related expectations. For example, in a review of measurements for expectations of patients with musculoskeletal disorders (van Hartingsveld et al., 2010), only one out of 24 studies attempted to measure both features. Assessing these aspects of expectations separately could facilitate a more differential understanding of expectation effects and would help to inform the design of interventions targeted at patients' expectations in medical conditions. Of the instruments described above, only the IPQ-R and the B-IPQ offer the possibility to assess several aspects of expectations on distinct scales. An alternative option would be the parallel use of validated instruments for both treatment-related expectations and behavior-related expectations.

Another neglected aspect is the assessment of patients' expectations regarding adverse effects or side effects of treatment and health behavior. As described above, few instruments assess side-effect expectations. While some measurement instruments incorporate both items about positive and negative outcome expectations (see **Table 1**), they are often subsumed in one scale (by reverse-coding items with negative expectations). However, expectations about positive and negative effects do not necessarily belong in one dimension. As an example, a study assessing expectations of patients who had undergone implantable cardioverter defibrillator implantation (Habibovic et al., 2014) revealed two distinguishable factors of positive and negative expectations, of which only negative expectations predicted higher levels of anxiety, depression and concerns at 3-month follow-up. Distinguishing between expectations of benefits and adverse effects might be especially valuable if they affect different dimensions of outcome and different timeframes. For instance, a patient undergoing coronary artery bypass graft surgery might expect a benefit in reducing shortness of breath in the long term, but might also expect pain in the short-term post-surgery period. In such a scenario, summing up the two aspects of expectations would be counterintuitive. While the majority of existing measurement instruments assess benefit expectation, only a small number have been used to separately assess side-effect expectations. Moreover, we are not aware of any instrument assessing expected adverse effects of health behaviors. Assessing these side effects might explain additional variance in patients engaging or not engaging in health-related behavior.

Further aspects that are underrepresented in studies assessing expectations are stimulus/external outcome expectations, process/structural expectations and timeline expectations. As mentioned above, outcome expectations can be related to internal response expectations or to expectations regarding external effects of illness and treatment, such as financial consequences or consequences affecting significant others. The majority of measurement instruments, however, focus on response expectancies. External factors, such as the consequences of treatment on a spouse, can be of significant importance in patients undergoing medical treatment. Therefore, assessing such external outcome expectations might further complete the picture of patients' expectations.

Expectations about the process and the structure of treatments are more difficult to assess in complex treatments, which might be a reason why few instruments attempt to capture these aspects. Relevantly, evidence from qualitative research shows that patients do hold quite specific process- and structure-related expectations (Haanstra et al., 2013). As these aspects are related to treatment outcome (see above), it would be worthwhile to assess them more systematically in patients in medical care. Finally, expectations about the temporal course of a disease have been shown to be predictive of several health outcomes across medical conditions (Broadbent et al., 2015). So far, this aspect of expectations has most often been operationalized explicitly in studies using the IPQ-R and B-IPQ. Given their predictive value, future studies should consider assessing expectations regarding temporal course more often.

Specific vs. Generalized Assessment of Expectations

As expectations are to a substantial extent situation-specific, the majority of instruments assess expectations for a specific treatment of a particular medical condition. As a result, even within one single category of medical conditions (e.g., musculoskeletal; van Hartingsveld et al., 2010; Zywiel et al., 2013), a high heterogeneity of expectation assessment can be found. This makes it difficult to compare the differential impact of certain expectations across different treatments and illnesses.

Likewise, with regard to the assessment of outcome expectations too, instruments differ in their specificity, assessing expectations about rather disease-specific symptoms or functions (e.g., degree of joint rotation, sexual functioning), generic symptoms (e.g., pain, sleep), broadly applicable concepts like disability, quality of life or return to work, or trait-like generalized outcome expectations (e.g., optimism, hope). Many instruments assess expectations on a disease- or treatment-specific level, meaning that they are not applicable to other conditions. Thus, expectation effects cannot be compared across conditions. The assessment of generalized outcome expectations like optimism is possible for any condition. However, this does not provide any insight into the patient's expectations while receiving medical treatment, as such instruments capture expectations on a very abstract level, with no specific reference to the treatment context. A solution to balance these two goals might be to measure expectations regarding expected disability, quality of life, or return to work (see **Table 1**). In contrast to disease-specific outcome instruments, the assessment of these kinds of expectations would be applicable to any disease or treatment. At the same time, such an assessment could still ask about concrete entities that are relevant for the patient's specific illness and treatment experience, as opposed to assessing outcome expectations on a very abstract basis, as is the case with optimism and similar concepts.

Additional Aspects to Consider

In addition to the aforementioned points, the timing of the assessment should be taken into consideration when assessing patients' expectations: Expectations have been assessed before, shortly after or at recovery/follow-up of a treatment or diagnostic test (Zywiel et al., 2013). Most studies have assessed expectations prior to the treatment or the diagnostic procedure (van Hartingsveld et al., 2010), which seems logical since these are salient events that are likely to trigger expectations. Presumably, expectations might be influenced by the course of treatment or diagnostic procedure. However, the effects of different assessment timing remain unclear, as they have rarely been investigated systematically (e.g., van den Akker-Scheek et al., 2007). Therefore, to investigate the temporal course of patients' expectations and the influencing factors, they should be assessed at multiple time points in the course of a treatment or a diagnostic procedure (Kamper et al., 2015). Moreover, assessing expectations on multiple occasions (before, during, and after a procedure) might foster knowledge about the stability of expectations. Additionally, researchers should always consider the burden of assessment with regard to the patient's condition. However, as most expectation scales are brief and intuitive, this

should not be a problem in most cases. Finally, although the main focus of this review was on patients' expectations, the expectations of healthcare providers/physicians may also play a critical role for treatment outcomes. Studies examining the relevance of physicians' expectations are scarce, although they have been shown to be related to treatment outcomes at least in some studies (e.g., Gracely et al., 1985; Galer et al., 1997; Witt et al., 2012). Further, there is evidence that if physicians communicate their high expectations to their patients, the patients' expectations are increased (Crow et al., 1999; Verheul et al., 2010). Certainly, it could be valuable to assess physicians' expectations and their impact on treatment outcomes in order to further explore the role of expectations in the medical treatment context. In particular, future studies should endeavor to elucidate the relationship between physicians' expectations and patients' expectations. The latter may mediate the effects of the former on treatment outcomes.

CONCLUSION

Patients' expectations in the context of medical treatment constitute a promising area of research, as growing evidence suggests that they have an influence on health outcomes across a variety of medical conditions. However, the aggregation of evidence is complicated by an inconsistent and disintegrated application of expectation constructs and the heterogeneity of assessment strategies. Within this review, we outlined an integrative model of expectations that aims to facilitate the consistent use of expectation constructs and more theory-driven standardized assessment strategies. In particular, the application of standardized, psychometrically evaluated measures, assessing multidimensional aspects of patients' expectations that are applicable across various medical treatments has the potential to generate a more comprehensive understanding of expectation effects in medical treatments. Future research should overcome the current obstacles in assessing expectations as outlined above. Moreover, more research is needed on the interrelation of different expectation aspects as well as on factors influencing patients' expectations of illness and treatment in clinical populations. Most studies investigating this question in medical patients have done so cross-sectionally (e.g., Scott et al., 2012; Laferton et al., 2015a). Prospective studies are warranted to gain a better understanding of the direction of influencing variables (e.g., demographic, medical, and psychosocial).

This might ultimately facilitate interventions aiming to influence patients' expectations in order to improve health outcomes. Patients' expectations can be effectively modulated by verbally suggesting that treatment is beneficial (Bingel et al., 2011; Kam-Hansen et al., 2014), using an empathetic interaction style (Kaptchuk et al., 2008), or discussing patients' treatment beliefs and concepts (Laferton et al., 2015b). Recently, several clinical intervention studies have shown that patients' expectations can be optimized via brief psychological interventions and that these interventions ultimately lead to improved health outcomes (Broadbent et al., 2009; von Blanckenburg et al., 2013, 2015; Rief et al., 2017). The application of theory guided frameworks, such as the ViolEx-model on expectation development, expectation maintenance, and expectation change proposed by Rief and Petrie (2016), might further help to refine such interventions. In this regard, an integrated understanding and assessment of patients' expectations is the first step toward improved health care across medical conditions.

AUTHOR CONTRIBUTIONS

JL, TK, CA, SS, and MS-M: Substantial contributions to the conception and design of the manuscript; drafting the work or revising it critically for important intellectual content; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This review was not supported by funding of any sort.

ACKNOWLEDGMENT

The authors thank Angelika Weigel and Thomas Munder for their helpful comments on this manuscript.

REFERENCES

Armitage, C. J., Norman, P., Alganem, S., and Conner, M. (2015). Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. *Ann. Behav. Med.* 49, 239–246. doi: 10.1007/s12160-014-9653-4

Auer, C. J., Glombiewski, J. A., Doering, B. K., Winkler, A., Laferton, J. A. C., Broadbent, E., et al. (2016a). Patients' expectations predict surgery outcomes: a meta-analysis. *Int. J. Behav. Med.* 23, 49–62. doi: 10.1007/s12529-015-9500-4

Auer, C. J., Kube, T., Laferton, J. A. C., Salzmann, S., Shedden-Mora, M. C., Rief, W., et al. (2016b). Welche erwartungen sagen postoperative depressivität und ängstlichkeit bei herzchirurgischen patienten am stärksten vorher? [Which expectations have the strongest predictive value for postsurgical depression and anxiety in patients undergoing CABG surgery]. *Z. Klin. Psychol. Psychother.* 45, 93–108. doi: 10.1026/1616-3443/a000358

Axelrad, K. J. (1982). Locus of control and causal attributions as they relate to expectations for coping with a heart attack. *Diss. Abstr. Int.* 42:4924.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York, NY: W. H. Freeman and Company.

Bandura, A. (2001). Social cognitive theory: an agentic perspective. *Annu. Rev. Psychol.* 52, 1–26. doi: 10.1146/annurev.psych.52.1.1

Barefoot, J. C., Brummett, B. H., Williams, R. B., Siegler, I. C., Helms, M. J., Boyle, S. H., et al. (2011). Recovery expectations and long-term prognosis of patients with coronary heart disease. *Arch. Intern. Med.* 171, 929–935. doi: 10.1001/archinternmed.2011.41

Barsky, A. J., Saintfort, R., Rogers, M. P., and Borus, J. F. (2002). Nonspecific medication side effects and the nocebo phenomenon. *JAMA* 287, 622–627.

Becker, M. H. ed (1974). *The Health Belief Model and Personal Health Behavior*. Thorofare, NJ: Slack.

Bingel, U., Wanigasekera, V., Wiech, K., Ni Mhuircheartaigh, R., Lee, M. C., Ploner, M., et al. (2011). The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid remifentanil. *Sci. Transl. Med.* 3:70ra14. doi: 10.1126/scitranslmed.3001244

Booth-Kewley, S., Schmied, E. A., Highfill-McRoy, R. M., Sander, T. C., Blivin, S. J., and Garland, C. F. (2014). A prospective study of factors affecting recovery from musculoskeletal injuries. *J. Occup. Rehabil.* 24, 287–296. doi: 10.1007/s10926-013-9456-7

Bowling, A., Rowe, G., Lambert, N., Waddington, M., Mahtani, K. R., Kenten, C., et al. (2012). The measurement of patients' expectations for health care: a review and psychometric testing of a measure of patients' expectations. *Health Technol. Assess.* 16, i–xii. doi: 10.3310/hta16300 , 1-509,

Broadbent, E., Ellis, C. J., Thomas, J., Gamble, G., and Petrie, K. J. (2009). Further development of an illness perception intervention for myocardial infarction patients: a randomized controlled trial. *J. Psychosom. Res.* 67, 17–23. doi: 10.1016/j.jpsychores.2008.12.001

Broadbent, E., Petrie, K. J., Main, J., and Weinman, J. (2006). The brief illness perception questionnaire. *J. Psychosom. Res.* 60, 631–637. doi: 10.1016/j.jpsychores.2005.10.020

Broadbent, E., Wilkes, C., Koschwanez, H., Weinman, J., Norton, S., and Petrie, K. J. (2015). A systematic review and meta-analysis of the brief illness perception questionnaire. *Psychol. Health* 30, 1361–1385. doi: 10.1080/08870446.2015.1070851

Brummett, B. H., Babyak, M. A., Mark, D. B., Clapp-Channing, N. E., Siegler, I. C., and Barefoot, J. C. (2004). Prospective study of perceived stress in cardiac patients. *Ann. Behav. Med.* 27, 22–30.

Cameron, L. D., and Leventhal, H. (eds) (2003). *The Self-Regulation of Health and Illness Behavior*. New York, NY: Routledge.

Carver, C. S., Scheier, M. F., and Segerstrom, S. C. (2010). Optimism. *Clin. Psychol. Rev.* 30, 879–889. doi: 10.1016/j.cpr.2010.01.006

Chunta, K. S. (2009). Expectations, anxiety, depression, and physical health status as predictors of recovery in open-heart surgery patients. *J. Cardiovasc. Nurs.* 24, 454–464. doi: 10.1097/JCN.0b013e3181ac8a3c

Colagiuri, B., and Zachariae, R. (2010). Patient expectancy and post-chemotherapy nausea: a meta-analysis. *Ann. Behav. Med.* 40, 3–14. doi: 10.1007/s12160-010-9186-4

Colloca, L., and Finniss, D. (2012). Nocebo effects, patient-clinician communication, and therapeutic outcomes. *JAMA* 307, 567–568. doi: 10.1001/jama.2012.115

Colloca, L., and Miller, F. G. (2011a). How placebo responses are formed: a learning perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 1859–1869. doi: 10.1098/rstb.2010.0398

Colloca, L., and Miller, F. G. (2011b). Role of expectations in health. *Curr. Opin. Psychiatry* 24, 149–155. doi: 10.1097/YCO.0b013e328343803b

Constantino, M. J., Arnkoff, D. B., Glass, C. R., Ametrano, R. M., and Smith, J. Z. (2011). Expectations. *J. Clin. Psychol.* 67, 184–192. doi: 10.1002/jclp.20754

Crane, M. M., Ward, D. S., Lutes, L. D., Bowling, J. M., and Tate, D. F. (2016). Theoretical and behavioral mediators of a weight loss intervention for men. *Ann. Behav. Med.* 50, 460–470. doi: 10.1007/s12160-016-9774-z

Crow, R., Gage, H., and Hampson, S. (1999). The role of expectancies in the placebo effect and their use in the delivery of health care a systematic review. *Health Technol. Assess.* 3, 1–96.

David, D., Montgomery, G. H., Stan, R., DiLorenzo, T., and Erblich, J. (2004). Discrimination between hopes and expectancies for nonvolitional outcomes: psychological phenomenon or artifact? *Pers. Individ. Dif.* 36, 1945–1952. doi: 10.1016/j.paid.2003.08.013

de Craen, A. J. M., Tijssen, J. G. P., de Gans, J., and Kleijnen, J. (2000). Placebo effect in the acute treatment of migraine: subcutaneous placebos are better than oral placebos. *J. Neurol.* 247, 183–188. doi: 10.1007/s004150050560

Devilly, G. J., and Borkovec, T. D. (2000). Psychometric properties of the credibility/expectancy questionnaire. *J. Behav. Ther. Exp. Psychiatry* 31, 73–86.

Di Blasi, Z., Harkness, E., Ernst, E., Georgiou, A., and Kleijnen, J. (2001). Influence of context effects on health outcomes: a systematic review. *Lancet* 357, 757–762.

Dohnke, B., Müller-Fahrnow, W., and Knäuper, B. (2006). Der einfluss von ergebnis- und selbstwirksamkeitserwartungen auf die ergebnisse einer rehabilitation nach hüftgelenkersatz [The influence of outcome- and self-efficacy expectations on rehabilitation results after hip joint replacement]. *Z. Gesundheitspsychol.* 14, 11–20. doi: 10.1026/0943-8149.14.1.11

Dougherty, C. M., Johnston, S. K., and Thompson, E. A. (2007). Reliability and validity of the self-efficacy expectations and outcome expectations after implantable cardioverter defibrillator implantation scales. *Appl. Nurs. Res.* 20, 116–124. doi: 10.1016/j.apnr.2007.04.004

Enck, P., Bingel, U., Schedlowski, M., and Rief, W. (2013). The placebo response in medicine: minimize, maximize or personalize? *Nat. Rev. Drug Discov.* 12, 191–204. doi: 10.1038/nrd3923

Fadyl, J., and McPherson, K. (2008). Return to work after injury: a review of evidence regarding expectations and injury perceptions, and their influence on outcome. *J. Occup. Rehabil.* 18, 362–374. doi: 10.1007/s10926-008-9153-0

Ferrari, R., Vidotto, G., Ferraro, T., Tosato, F., and Milite, D. (2015). Recovery expectations and quality of life after revascularization treatments. *Br. J. Med. Med. Res.* 5, 1431–1440. doi: 10.9734/BJMMR/2015/14264

Francis, J. J., Wileman, S. M., Bekker, H., Barton, G. R., and Ramsay, C. R. (2009). Beliefs about surgery: development and validation of an instrument to assess cognitive representations about surgery, in the context of a chronic illness. *Psychol. Health* 24, 1125–1137. doi: 10.1080/08870440802071157

Galer, B. S., Schwartz, L., and Turner, J. A. (1997). Do patient and physician expectations predict response to pain-relieving pro-cedures? *Clin. J. Pain* 13, 348–351.

Geers, A. L., Wellman, J. A., Fowler, S. L., Helfer, S. G., and France, C. R. (2010). Dispositional optimism predicts placebo analgesia. *J. Pain* 11, 1165–1171. doi: 10.1016/j.jpain.2010.02.014

Glaesmer, H., Rief, W., Martin, A., Zenger, M., Hinz, A., and Br, E. (2012). Psychometric properties and population-based norms of the Life Orientation Test Revised (LOT-R). *Br. J. Health Psychol.* 17, 432–445. doi: 10.1111/j.2044-8287.2011.02046.x

Gracely, R., Dubner, R., Deeter, W., and Wolskee, P. (1985). Clinicians' expectations influence placebo analgesia. *Lancet* 1:43. doi: 10.1016/S0140-6736(85)90984-5

Haanstra, T. M., Hanson, L., Evans, R., van Nes, F. A., De Vet, H. C. W., Cuijpers, P., et al. (2013). How do low back pain patients conceptualize their expectations regarding treatment? Content analysis of interviews. *Eur. Spine J.* 22, 1986–1995. doi: 10.1007/s00586-013-2803-8

Haanstra, T. M., Kamper, S. J., Williams, C. W., Spriensma, A. S., Lin, C.-W. C., Maher, C. G., et al. (2015a). Does adherence to treatment mediate the relationship between patients' treatment outcome expectancies and the outcomes pain intensity and recovery from acute low back pain? *Pain* 156, 1530–1539. doi: 10.1097/j.pain.0000000000000198

Haanstra, T. M., Tilbury, C., Kamper, S. J., Tordoir, R. L., Vliet Vlieland, T. P. M., Nelissen, R. G. H. H., et al. (2015b). Can optimism, pessimism, hope, treatment credibility and treatment expectancy be distinguished in patients undergoing total hip and total knee arthroplasty? *PLoS ONE* 10:e0133730. doi: 10.1371/journal.pone.0133730

Haanstra, T. M., van den Berg, T., Ostelo, R. W., Poolman, R. W., Jansma, E. P., Jansma, I. P., et al. (2012). Systematic review: do patient expectations influence treatment outcomes in total knee and total hip arthroplasty? *Health Qual. Life Outcomes* 10:152. doi: 10.1186/1477-7525-10-152

Habibovic, M., Pedersen, S. S., van den Broek, K. C., and Denollet, J. (2014). Monitoring treatment expectations in patients with an implantable cardioverter-defibrillator using the EXPECT-ICD scale. *Europace* 16, 1022–1027. doi: 10.1093/europace/euu006

Hagger, M., and Orbell, S. (2003). A meta-analytic review of the common-sense model of illness representations. *Psychol. Health* 18, 141–184. doi: 10.1080/088704403100081321

Hanssen, M. M., Peters, M. L., Vlaeyen, J. W. S., Meevissen, Y. M. C., and Vancleef, L. M. G. (2013). Optimism lowers pain: evidence of the causal status and underlying mechanisms. *Pain* 154, 53–58. doi: 10.1016/j.pain.2012.08.006

Heisig, S. R., Shedden-Mora, M. C., Hidalgo, P., and Nestoriuc, Y. (2015). Framing and personalizing informed consent to prevent negative expectations: an experimental pilot study. *Health Psychol.* 34, 1033–1037. doi: 10.1037/hea0000217

Heisig, S. R., Shedden-Mora, M. C., von Blanckenburg, P., Rief, W., Witzel, I., Albert, U.-S., et al. (2016). What do women with breast cancer expect from

their treatment? Correlates of negative treatment expectations about endocrine therapy. *Psychooncology* 25, 1485–1492. doi: 10.1002/pon.4089

Hirani, S. P., Patterson, D. L. H., and Newman, S. P. (2008). What do coronary artery disease patients think about their treatments? An assessment of patients' treatment representations. *J. Health Psychol.* 13, 311–322. doi: 10.1177/1359105307088133

Holden, G. (1991). Social work in health care the relationship of self- efficacy appraisals to subsequent health related outcomes. *Soc. Work Health Care* 16, 53–93. doi: 10.1300/J010v16n01_05

Holmes, S. D., Fornaresio, L. M., Miller, C. E., Shuman, D. J., and Ad, N. (2016). Development of the Cardiac Surgery Patient Expectations Questionnaire (C-SPEQ). *Qual. Life Res.* 25, 2077–2086. doi: 10.1007/s11136-016-1243-4

Horne, R., Weinman, J., and Hankins, M. (1999). The beliefs about medicines questionnaire: the development and evaluation of a new method for assessing the cognitive representation of medication. *Psychol. Health* 14, 1–24. doi: 10.1080/08870449908407311

Hüppe, M., Beckhoff, M., Klotz, K.-F., Heinzinger, M., Prüßmann, M., Gerlach, K., et al. (2003). Reliabilität und validität des anästhesiologischen nachbefragungsbogens bei elektiv operierten patienten [Reliability and validity of the anaesthesilogical questionnaire for electively operated patients]. *Anaesthesist* 52, 311–320. doi: 10.1007/s00101-003-0471-5

Hüppe, M., Kemter, A., Schmidtke, C., and Klotz, K.-F. (2013). Postoperative beschwerden: geschlechtsunterschiede in erwartung, auftreten und bewertung [Postoperative complaints: gender differences in expectations, prevalence, and appraisal]. *Anaesthesist* 62, 528–536. doi: 10.1007/s00101-013-2182-x

Jenkins, L. S., and Gortner, S. R. (1998). Correlates of self-efficacy expectation and prediction of walking behavior in cardiac surgery elders. *Ann. Behav. Med.* 20, 99–103. doi: 10.1007/BF02884455

Johannessen, K. B., Oettingen, G., and Mayer, D. (2012). Mental contrasting of a dieting wish improves self-reported health behaviour. *Psychol. Health* 27(Suppl. 2), 43–58. doi: 10.1080/08870446.2011.626038

Jones, F., and Riazi, A. (2011). Self-efficacy and self-management after stroke: a systematic review. *Disabil. Rehabil.* 33, 797–810. doi: 10.3109/09638288.2010.511415

Juergens, M. C., Seekatz, B., Moosdorf, R. G., Petrie, K. J., and Rief, W. (2010). Illness beliefs before cardiac surgery predict disability, quality of life, and depression 3 months later. *J. Psychosom. Res.* 68, 553–560. doi: 10.1016/j.jpsychores.2009.10.004

Kam-Hansen, S., Jakubowski, M., Kelley, J. M., Kirsch, I., Hoaglin, D. C., Kaptchuk, T. J., et al. (2014). Altered placebo and drug labeling changes the outcome of episodic migraine attacks. *Sci. Transl. Med.* 6:218ra5. doi: 10.1126/scitranslmed.3006175

Kamper, S. J., Kongsted, A., Haanstra, T. M., and Hestbaek, L. (2015). Do recovery expectations change over time? *Eur. Spine J.* 24, 218–226. doi: 10.1007/s00586-014-3380-1

Kappes, H. B., and Oettingen, G. (2011). Positive fantasies about idealized futures sap energy. *J. Exp. Soc. Psychol.* 47, 719–729. doi: 10.1016/j.jesp.2011.02.003

Kaptchuk, T. J., Kelley, J. M., Conboy, L. A., Davis, R. B., Kerr, C. E., Jacobson, E. E., et al. (2008). Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. *BMJ* 336, 999–1003. doi: 10.1136/bmj.39524.439618.25

Kirsch, I. (1983). Response expectancy as a determinant of experience and behavior. *Am. Psychol.* 40, 1189–1202.

Kirsch, I. (1995). "Self-efficacy and outcome expectancies: a concluding commentary," in *Self-Efficacy, Adaption, and Adjustment: Theory, Research and Application*, ed. J. E. Maddux (New York, NY: Plenum Press), 331–345.

Kirsch, I. (1997). Response expectancy theory and application: a decennial review. *Appl. Prev. Psychol.* 6, 69–79. doi: 10.1016/S0962-1849(05)80012-5

Kirsch, I. (2004). Conditioning, expectancy, and the placebo effect: comment on Stewart-Williams and Podd (2004). *Psychol. Bull.* 130, 341–343; discussion 344–345. doi: 10.1037/0033-2909.130.2.341

Kirsch, I., Kong, J., Sadler, P., Spaeth, R., Cook, A., Kaptchuk, T. J., et al. (2014). Expectancy and conditioning in placebo analgesia: Separate or connected processes? *Psychol. Conscious.* 1, 51–59. doi: 10.1037/cns0000007

Kirsch, I., Lynn, S. J., Vigorito, M., and Miller, R. R. (2004). The role of cognition in classical and operant conditioning. *J. Clin. Psychol.* 60, 369–392. doi: 10.1002/jclp.10251

Kravitz, R. L. (1996). Patients' expectations for medical care: an expanded formulation based on review of the literature. *Med. Care Res. Rev.* 53, 3–27. doi: 10.1177/107755879605300101

Kube, T., D'Astolfo, L., Glombiewski, J. A., Doering, B. K., and Rief, W. (2016). Focusing on situation-specific expectations in major depression as basis for behavioral experiments - development of the Depressive Expectations Scale (DES). *Psychol. Psychother.* doi: 10.1111/papt.12114 [Epub ahead of print].

Kube, T., Rief, W., and Glombiewski, J. A. (2017). On the maintenance of expectations in major depression – investigating a neglected phenomenon. *Front. Psychol.* 8:9. doi: 10.3389/fpsyg.2017.00009

Laferton, J. A. C., Auer, C. J., Shedden-Mora, M. C., Moosdorf, R., and Rief, W. (2015a). Factors associated with disability expectations in patients undergoing heart surgery. *Int. J. Behav. Med.* 22, 85–91. doi: 10.1007/s12529-014-9434-2

Laferton, J. A. C., Auer, C. J., Shedden-Mora, M. C., Moosdorf, R., and Rief, W. (2015b). Optimizing preoperative expectations in cardiac surgery patients is moderated by level of disability: the successful development of a brief psychological intervention. *Psychol. Health Med.* 21, 272–285. doi: 10.1080/13548506.2015.1051063

Laferton, J. A. C., Shedden Mora, M., Auer, C. J., Moosdorf, R., and Rief, W. (2013). Enhancing the efficacy of heart surgery by optimizing patients' preoperative expectations: study protocol of a randomized controlled trial. *Am. Heart J.* 165, 1–7. doi: 10.1016/j.ahj.2012.10.007

Leedham, B., Meyerowitz, B. E., Muirhead, J., and Frist, W. H. (1995). Positive expectations predict health after heart transplantation. *Health Psychol.* 14, 74–79. doi: 10.1037/0278-6133.14.1.74

Leung, K. K., Silvius, J. L., Pimlott, N., Dalziel, W., and Drummond, N. (2009). Why health expectations and hopes are different: the development of a conceptual model. *Health Expect.* 12, 347–360. doi: 10.1111/j.1369-7625.2009.00570.x

Leventhal, H., Meyer, D., and Nerenz, D. R. (1980). "The common sense representation of illness danger," in *Contributions to Medical Psychology*, ed. S. Rachman (New York, NY: Pergamon Press), 17–30.

Maddux, J. E. (1999). "Expectancies and the social-cognitive-perspective: basic principles, processes, and variables," in *How Expectancies Shape Experience*, ed. I. Kirsch (Washington, DC: American Psychological Association), 17–39.

Maddux, J. E. (2007). "Expectations and health," in *Cambridge Handbook of Psychology, Health and Medicine*, eds S. Ayers, A. Baum, C. McManus, S. Newmann, K. Wallston, J. Weinmann, et al. (Cambridge: Cambridge University Press), 87–92.

Mahomed, N. N., Liang, M. H., Cook, E. F., Daltroy, L. H., Fortin, P. R., Fossel, A. H., et al. (2002). The importance of patient expectations in predicting functional outcomes after total joint arthroplasty. *J. Rheumatol.* 29, 1273–1279.

McCarthy, S., Lyons, A., Weinman, J., Talbot, R., and Purnell, D. (2003). Do expectations influence recovery from oral surgery? An illness representation approach. *Psychol. Health* 18, 109–126. doi: 10.1080/0887044031000080674

Mondloch, M. V., Cole, D. C., and Frank, J. W. (2001). Does how you do depend on how you think you'll do? A systematic review of the evidence for a relation between patients' recovery expectations and health outcomes. *CMAJ* 165, 174–179.

Moser, D. K., and Dracup, K. (1995). Psychosocial recovery from a cardiac event: the influence of perceived control. *Heart Lung* 24, 273–280. doi: 10.1016/S0147-9563(05)80070-6

Moser, D. K., Riegel, B., Mckinley, S., Doering, L. V., Meischke, H., Heo, S., et al. (2009). The control attitudes scale-revised. *Nurs. Res.* 58, 42–51. doi: 10.1097/NNR.0b013e3181900ca0

Moss-Morris, R., Weinman, J., Petrie, K. J., Horne, R., Cameron, L., and Buick, D. (2002). The revised illness perception questionnaire (IPQ-R). *Psychol. Health* 17, 1–16. doi: 10.1080/08870440290001494

Murgatroyd, D. F., Harris, I. A., Tran, Y., Cameron, I. D., Gabbe, B., Cameron, P., et al. (2016). Predictors of return to work following motor vehicle related orthopaedic trauma. *BMC Musculoskelet. Disord.* 17:171. doi: 10.1186/s12891-016-1019-6

Nestoriuc, Y., von Blanckenburg, P., Schuricht, F., Barsky, A. J., Hadji, P., Albert, U.-S., et al. (2016). Is it best to expect the worst? Influence of patients' side-effect expectations on endocrine treatment outcome in a 2-year prospective clinical cohort study. *Ann. Oncol.* 27, 1909–1915. doi: 10.1093/annonc/mdw266

Noble, P. C., Scuderi, G. R., Brekke, A. C., Sikorskii, A., Benjamin, J. B., Lonner, J. H., et al. (2012). Development of a new knee society scoring system. *Clin. Orthop. Relat. Res.* 470, 20–32. doi: 10.1007/s11999-011-2152-z

Oettingen, G. (2012). Future thought and behaviour change. *Eur. Rev. Soc. Psychol.* 23, 1–63. doi: 10.1080/10463283.2011.643698

Oettingen, G., and Mayer, D. (2002). The motivating function of thinking about the future: expectations versus fantasies. *J. Pers. Soc. Psychol.* 83, 1198–1212. doi: 10.1037//0022-3514.83.5.1198

Oettingen, G., and Wadden, T. A. (1991). Expectation, fantasy, and weight loss: is the impact of positive thinking always positive? *Cognit. Ther. Res.* 15, 167–175. doi: 10.1007/BF01173206

Petrie, K. J., Jago, L. A., and Devcich, D. A. (2007). The role of illness perceptions in patients with medical conditions. *Curr. Opin. Psychiatry* 20, 163–167.

Petrie, K. J., and Weinman, J. (2012). Patients' perceptions of their illness the dynamo of volition in health care. *Curr. Dir. Psychol. Sci.* 21, 60–65. doi: 10.1177/0963721411429456

Petrie, K. J., Weinman, J., Sharpe, N., and Buckley, J. (1996). Role of patients' view of their illness in predicting return to work and functioning after myocardial infarction: longitudinal study. *Br. Med. J.* 312, 1191–1194. doi: 10.1136/bmj.312.7040.1191

Powell, R., Johnston, M., Smith, W. C., King, P. M., Chambers, W. A., Krukowski, Z., et al. (2012). Psychological risk factors for chronic post-surgical pain after inguinal hernia repair surgery: a prospective cohort study. *Eur. J. Pain* 16, 600–610. doi: 10.1016/j.ejpain.2011.08.010

Price, D. D., Finniss, D. G., and Benedetti, F. (2008). A comprehensive review of the placebo effect: recent advances and current thought. *Annu. Rev. Psychol.* 59, 565–590. doi: 10.1146/annurev.psych.59.113006.095941

Rief, W., Barsky, A. J., Glombiewski, J. A., Nestoriuc, Y., and Glaesmer, H. (2011). Assessing general side effects in clinical trials: reference data from the general population. *Pharmacoepidemiol. Drug Saf.* 20, 405–415. doi: 10.1002/pds

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., and Thorwart, A. (2015). Expectancies as core features of mental disorders. *Curr. Opin. Psychiatry* 28, 378–385. doi: 10.1097/YCO.0000000000000184

Rief, W., and Petrie, K. J. (2016). Can psychological expectation models be adapted for placebo research? *Front. Psychol.* 7:1876. doi: 10.3389/fpsyg.2016.01876

Rief, W., Shedden-Mora, M. C., Laferton, J. A. C., Auer, C. J., Petrie, K. J., Salzmann, S., et al. (2017). Preoperative optimization of patient expectations improves long-term outcome in heart surgery patients: results of the randomized controlled PSY-HEART trial. *BMC Med.* 15:4. doi: 10.1186/s12916-016-0767-3

Scheier, M. F., and Carver, C. S. (1985). Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health Psychol.* 4, 219–247.

Scheier, M. F., Carver, C. S., and Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. *J. Pers. Soc. Psychol.* 67, 1063–1078.

Schwarzer, R. (1994). Optimism, vulnerability, and self-beliefs as health-related cognitions: a systematic overview. *Psychol. Health* 9, 161–180. doi: 10.1080/08870449408407475

Schwarzer, R., and Jerusalem, M. (1995). "Generalized Self-Efficacy scale," in *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*, eds J. Weinman, S. Wright, and M. Johnston (Windsor: NFER-NELSON), 35–37.

Schwarzer, R., Luszczynska, A., Ziegelmann, J. P., Scholz, U., and Lippke, S. (2008). Social-cognitive predictors of physical exercise adherence: three longitudinal studies in rehabilitation. *Health Psychol.* 27, S54–S63. doi: 10.1037/0278-6133.27.1(Suppl.).S54

Schwarzer, R., and Renner, B. (2016). *Health-Specific Self-Efficacy Scales.* 1–21. Available at: http://userpage.fu-berlin.de/%7Ehealth/healself.pdf [Accessed October 5, 2016].

Scott, C. E. H., Bugler, K. E., Clement, N. D., MacDonald, D., Howie, C. R., and Biant, L. C. (2012). Patient expectations of arthroplasty of the hip and knee. *J. Bone Joint Surg. Br.* 94, 974–981. doi: 10.1302/0301-620X.94B7.28219

Sears, S. F., Serber, E. R., Lewis, T. S., Walker, R. L., Conners, N., Lee, J. T., et al. (2004). Do positive health expectations and optimism relate to quality-of-life outcomes for the patient with an implantable cardioverter defibrillator? *J. Cardiopulm. Rehabil.* 24, 324–331.

Sokol, M. C., McGuigan, K. A., Verbrugge, R. R., and Epstein, R. S. (2005). Impact of medication adherence on hospitalization risk and healthcare cost. *Med. Care* 43, 521–530. doi: 10.1097/01.mlr.0000163641.86870.af

Tait, R. C., Chibnall, J. T., and Krause, S. (1990). The pain disability index: psychometric properties. *Pain* 40, 171–182.

Tashjian, R. Z., Bradley, M. P., Tocci, S., Rey, J., Henn, R. F., and Green, A. (2007). Factors influencing patient satisfaction after rotator cuff repair. *J. Shoulder Elbow Surg.* 16, 752–758. doi: 10.1016/j.jse.2007.02.136

van den Akker-Scheek, I., Stevens, M., Groothoff, J. W., Bulstra, S. K., and Zijlstra, W. (2007). Preoperative or postoperative self-efficacy: which is a better predictor of outcome after total hip or knee arthroplasty? *Patient Educ. Couns.* 66, 92–99. doi: 10.1016/j.pec.2006.10.012

van Hartingsveld, F., Ostelo, R. W. J. G., Cuijpers, P., de Vos, R., Riphagen, I. I., and de Vet, H. C. W. (2010). Treatment-related and patient-related expectations of patients with musculoskeletal disorders: a systematic review of published measurement tools. *Clin. J. Pain* 26, 470–488. doi: 10.1097/AJP.0b013e3181e0ffd3

Verheul, W., Sanders, A., and Bensing, J. (2010). The effects of physicians' affect-oriented communication style and raising expectations on analogue patients' anxiety, affect and expectancies. *Patient Educ. Couns.* 80, 300–306. doi: 10.1016/j.pec.2010.06.017

von Blanckenburg, P., Schuricht, F., Albert, U.-S., Rief, W., and Nestoriuc, Y. (2013). Optimizing expectations to prevent side effects and enhance quality of life in breast cancer patients undergoing endocrine therapy: study protocol of a randomized controlled trial. *BMC Cancer* 13:426. doi: 10.1186/1471-2407-13-426

von Blanckenburg, P., Schuricht, F., Heisig, S. R., Shedden-Mora, M. C., Rehahn-Sommer, S., Albert, U.-S., et al. (2015). Psychologische optimierung von erwartungen zur prävention von nocebo-nebenwirkungen bei brustkrebs—2 fallberichte. *[Psychological optimization of expectations to prevent nocebo side effects in breast cancer—*2 case reports]. *Verhaltenstherapie* 25, 219–227. doi: 10.1159/000377711

Waldrop, D., Lightsey, O. R. Jr., Ethington, C. A., Woemmel, C. A., and Coke, A. L. (2001). Self-efficacy, optimism, health competence, and recovery from orthopedic surgery. *J. Couns. Psychol.* 48, 233–238. doi: 10.1037/0022-0167.48.2.233

Willett, W. C. (2002). Balancing life-style and genomics research for disease prevention. *Science* 296, 695–698. doi: 10.1126/science.1071055

Witt, C. M., Martins, F., Willich, S. N., and Schützler, L. (2012). Can i help you? Physicians' expectations as predictor for treatment outcome. *Eur. J. Pain* 16, 1455–1466. doi: 10.1002/j.1532-2149.2012.00152.x

Zywiel, M. G., Mahomed, A., Gandhi, R., Perruccio, A. V., and Mahomed, N. N. (2013). Measuring expectations in orthopaedic surgery: a systematic review. *Clin. Orthop. Relat. Res.* 471, 3446–3456. doi: 10.1007/s11999-013-3013-8

Conflict of Interest Statement: 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.

Copyright © 2017 Laferton, Kube, Salzmann, Auer and Shedden-Mora. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

OPINION
published: 06 February 2017
doi: 10.3389/fpsyg.2017.00112

Expectations of Social Inclusion and Exclusion

Eric D. Wesselmann[1]*, *James H. Wirth*[2] and *Michael J. Bernstein*[3]

[1] Department of Psychology, Illinois State University, Normal, IL, USA, [2] Department of Psychology, The Ohio State University at Newark, Newark, OH, USA, [3] Psychological and Social Science Program, Penn State University-Abington, Abington, PA, USA

Keywords: social exclusion, social inclusion, rejection, ostracism, sociometer theory, relational evaluation

OPEN ACCESS

Edited by:
Mario Gollwitzer,
University of Marburg, Germany

Reviewed by:
Jennifer Eck,
University of Mannheim, Germany
Selma Carolin Rudert,
University of Basel, Switzerland

***Correspondence:**
Eric D. Wesselmann
edwesse@ilstu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *29 October 2016*
Accepted: *16 January 2017*
Published: *06 February 2017*

Citation:
Wesselmann ED, Wirth JH and Bernstein MJ (2017) Expectations of Social Inclusion and Exclusion. Front. Psychol. 8:112. doi: 10.3389/fpsyg.2017.00112

Individuals engage in social interactions generally expecting inclusion (Kerr and Levine, 2008; Wesselmann et al., 2010, 2013). This expectation seems reasonable, given individuals' basic need to establish and maintain social connections to sustain physical and psychological well-being (Baumeister and Leary, 1995). However, individuals often experience *social exclusion*; situations broadly involving someone being disengaged or separated from others physically or emotionally (Riva and Eck, 2016). Exclusion experiences include various phenomena, such as interpersonal rejection, ostracism, and various types of discrimination (Smart Richman and Leary, 2009; Wesselmann et al., 2016).

These diverse threats to social inclusion are so detrimental, researchers argue that humans likely developed mechanisms safeguarding social inclusion (Lieberman, 2013), facilitating quick detection of threats to inclusionary status (Pickett and Gardner, 2005; Kerr and Levine, 2008; Wesselmann et al., 2012b). When threats occur, individuals experience cognitive and behavioral changes to facilitate recovery (Smart Richman and Leary, 2009; Williams, 2009). Considerable research has examined individuals' responses to exclusion, but less has focused on how expectations of inclusion or exclusion moderate those responses. In this article, we highlight research focused on how individuals calibrate their expectations of social inclusion and exclusion, and how these expectations influence the effect of exclusion on individuals' feelings of relational value and other adverse effects of social exclusion.

PERCEIVING EXCLUSIONARY CUES

Individuals monitor their environment for exclusionary cues using their *sociometer*, which detects fluctuations in an individual's relational evaluation (Leary, 1999; Leary and Baumeister, 2000). Relational evaluation is operationalized as "the degree to which others regard their relationship with the individual as valuable, important, or close" (Leary, 1999, p. 33). Individuals' perceived relational value is a proxy for inclusionary status (Leary, 1999; Leary and Baumeister, 2000). Exclusionary cues vary from direct to subtle (e.g., language, facial expression, non-verbal behaviors; Kerr and Levine, 2008), yet all produce feelings of social pain (Williams, 2009). Such exclusionary cues may be unambiguously clear, such as a partner stating they do not want to work with you (Maner et al., 2007), not being included during a game (Williams et al., 2000), or being treated in a cold and aloof manner (Geller et al., 1974; Wesselmann et al., 2010). Conversely, exclusion can occur in various subtle ways, such as not receiving eye contact from an avatar, which causes feelings of exclusion (Böckler et al., 2014) and lowered implicit self-esteem (Wirth et al., 2010). Even being stared *through* by a passerby (as if one does not exist) causes feelings of social disconnection (Wesselmann et al., 2012a).

Conversation dynamics can provide cues to one's inclusionary status (Koudenburg, 2014). Smooth conversations indicate relationship solidarity, while uncomfortable pauses are threatening to social connectedness (Koudenburg et al., 2011, 2013). Exclusion can occur during conversations when group members switch to a language unfamiliar to the target

(Hitlan et al., 2006; Dotan-Eliaz et al., 2009), when others use unknown acronyms (Hales et al., in press), through exclusive laughter (Klages and Wirth, 2014), or when the conversation makes people feel "out-of-the-loop"—when a person is included in the group, but feels excluded due to knowing there is information that they lack (Jones et al., 2009, 2011).

WHEN EXPECTATIONS OF INCLUSION ARE VIOLATED

During exclusion, relational devaluation occurs, and individuals suffer aversive physical and psychological consequences (Williams, 2009). However, little work has examined how *expectations* of inclusion/exclusion affect exclusion's consequences. Does expecting exclusion temper the negative outcomes, and unexpected exclusion intensify them, perhaps by threatening individuals' confidence in their sociometers? Because individuals monitor their environments for inclusion-relevant social cues (Leary, 1999; Williams, 2009), they likely experience unexpected exclusion more extremely than excepted exclusion. For example, Wesselmann et al. (2010) found that although excluded individuals aggressed more than included individuals, individuals who experienced *unexpected* exclusion demonstrated the most aggression and showed the least confidence in their sociometer. Further, Wirth et al. (2017) found participants who were unexpectedly excluded experienced increased basic need threat and negative affect, as well as decreased confidence in their sociometer, compared to participants who expected their exclusion. The latter group experienced need threat and negative affect once they received exclusionary social cues, and these negative effects continued on after they were ultimately excluded. Additionally, individuals who expected exclusion did not indicate decreased confidence in their sociometer between the time they received exclusionary cues and when they were excluded. Finally, Rudert and Greifeneder (2016) explicitly manipulated participants' expectations of situational norms and found that excluded participants experienced less negative effects when perceiving exclusion (rather than inclusion) as the norm. Collectively, these studies support neuroscience research suggesting that exclusion-related pain partially involves expectation violations (Somerville et al., 2006).

Based on previous theory (Leary, 1990; Wesselmann et al., 2016), relational evaluation is a key mechanism in understanding the degree to which social exclusion causes negative psychological outcomes. Specifically, deflated relational evaluation can cause negative feelings (Leary et al., 2001; Buckley et al., 2004) and may be related to lowered fulfillment of psychological needs, implicit self-esteem, and aggressive behavior temptations (Wirth et al., 2010; Bernstein et al., 2013). In response, individuals engage in behaviors aimed at safeguarding their relational evaluation. Socially excluded individuals have enhanced memory for social information (Gardner et al., 2000), increased desire to make new friends, and preferences for new potential interaction partners (Maner et al., 2007). Further, excluded individuals show increased attention to genuine signals of social inclusion (Bernstein et al., 2008, 2010a) and emotional expressions of happiness vs. anger (Sacco et al., 2011). Excluded individuals may be guided perceptually and behaviorally toward sources of social inclusion (i.e., increased attunement to positive, inclusive targets; DeWall et al., 2009). To our knowledge, no studies have directly assessed whether relational evaluation mediates excluded participants' perceptual and behavioral biases toward re-inclusion, but some evidence suggests participants' threatened need for belonging can mediate these effects (Bernstein et al., 2010a).

FUTURE RESEARCH QUESTIONS

Social Cue Attention and Response as Adaptation

Research on cognitive responses to exclusion are mixed: some studies show cognitive depletion (Baumeister et al., 2002), whereas others show cognitive benefits such as increased attention to and memory for social information (e.g., Gardner et al., 2000; Pickett et al., 2004; Bernstein et al., 2008). This apparent contradiction may be due to the paradigm used to examine the effects; paradigms revealing deficits tend to involve non-social tasks, while paradigms involving social tasks typically reveal benefits post-exclusion. Perhaps excluded individuals allocate available cognitive resources to tasks most effective to restoring relational evaluation levels, which non-social tasks may not do (Shilling and Brown, 2016).

To our knowledge, research has not directly tested this strategic re-distribution hypothesis (but see Gardner et al., 2000 comparing social and non-social memory). Such studies would strengthen the theoretical argument that individuals respond to exclusion in adaptive ways (i.e., survival-enhancing: Wesselmann et al., 2012b). Additionally, research should investigate if and how expectation violations influence any strategic re-distribution patterns. Williams (2009) argues that once individuals experience the immediate negative effects of exclusion, they subsequently focus cognitive resources on interpreting the situation to assess methods of recovery. These efforts may involve attributional processes or behavioral strategies. If individuals unexpectedly experience decreased relational evaluation, they may show strategic re-distribution more intensely than excluded individuals who expected exclusion because they experienced a more intense threat. Alternatively, simply experiencing *any* exclusion may trigger a strategic re-distribution response that is broad and undifferentiated to maximize re-inclusion efforts (Pickett and Gardner, 2005), and expectations may have little (or no) influence pattern.

Future research should examine if and how expectations matter for chronically excluded individuals. Williams (2009) refers to these individuals as being in the *resignation* stage, and argues that they likely come to expect exclusion in daily interactions. These individuals may experience learned helplessness that effectively comes from being unable to avoid exclusion or alter its consequences. Even though resigned individuals may anticipate exclusion, they may find unexpected exclusionary episodes (either in daily life or in a laboratory setting) more painful than other individuals precisely because they are caught unaware. Alternatively, resigned individuals may

simply be numb to the negative effects of exclusion regardless of their momentary expectations (Bernstein and Claypool, 2012; Riva et al., 2014).The resignation stage of exclusion is relatively new and research is sparse (but see Riva et al., 2016), so we can merely speculate on the influence of expectations in this context.

Paradigm Constraints, Expectation Cues, and Responses

Exclusion paradigms often blindside participants with exclusion (Williams and Wesselmann, 2011), but many exclusion experiences outside the laboratory likely involve some warning (Spoor and Williams, 2007; Kerr and Levine, 2008) or clear attributional information relevant during reflection (Nezlek et al., 2012). Thus, the conclusions drawn from most exclusion research may be limited in how well the research represents these everyday exclusion experiences. We have already discussed how Wirth et al. (2017) showed that the social cues prior to exclusion can affect individuals' expectations of, and ultimately responses to, their exclusion. Tuscherer et al. (2016) examined an additional understudied factor, participants' perceptions of fairness for an exclusion experience, and found that participants who perceived their exclusion as unfair experienced greater threat to efficacy needs (i.e., control and meaningful existence) than participants who perceived their exclusion as fair. These studies suggest that researchers should be mindful of how these two factors may relate to their future research questions and design their paradigms accordingly. Researchers should also consider how social cues and perceptions of fairness may influence each other both within a single exclusion episode and across subsequent episodes.

Participants' expectations of inclusion also likely influence *whether* participants will choose to respond pro- or anti-socially to exclusion, as well as the *degree* of their response, which is a current paradox in the literature (Wesselmann et al., 2015). Some research demonstrates that excluded participants will only respond pro-socially when they perceive the opportunity for re-affiliation (Maner et al., 2007; Mead et al., 2011). Potentially, any exclusion paradigm could be adapted to influence participants' expectations by manipulating explicit situational norms (Rudert and Greifeneder, 2016), confederate social cues in face-to-face or virtual *get-acquainted* paradigms (Wesselmann et al., 2010; Wirth et al., 2017), or explicit instructions involving opportunities to meet the target of participants' pro-/anti-social behavior (Maner et al., 2007). Researcher could also use the *life alone* paradigm (Twenge et al., 2001), which provides participants with fake feedback about their future social lives (e.g., their future will be lonely), as an expectations manipulation and then examine how those expectations influence the effects of subsequent exclusion using *in vivo* paradigms.

Further Integrating Relational Evaluation into Research

Researchers should investigate the specific role that relational evaluation plays in the consequences of exclusion, and how situation-level and individual-level characteristics influence this construct. *Situational factors* that may influence expectations of relational evaluation could be the psychological closeness of the sources of inclusion (e.g., a romantic partner; Arriaga et al., 2014), exclusion by in-group vs. out-group members (e.g., Gonsalkorale and Williams, 2007; Bernstein et al., 2010b; Goodwin et al., 2010; Cursan et al., 2016), or situations that require some type of exclusion (i.e., *role-based* exclusion, Nezlek et al., 2012; Rudert and Greifeneder, 2016); although exclusion in each case may hurt, violations of one's expected relational evaluation may help explain if and when exclusion may hurt more (or less) initially, and may also explain differential recovery (e.g., Wirth and Williams, 2009). *Individual factors* may also influence one's expected relational evaluation levels. For example, narcissistic individuals may expect high relational evaluation and thus respond with more aggression than non-narcissists when their expectations are violated (Bushman and Baumeister, 1998; Twenge and Campbell, 2003). Additionally, individuals high in rejection sensitivity (Downey and Feldman, 1996) may have lower expectations for perceived relational evaluation because they presume social interactions will not likely be positive. However, rejection-sensitive individuals expect exclusion in social situations, yet they respond with more hostility to exclusion than less-sensitive individuals (Ayduk et al., 2008; Pfundmair et al., 2015), suggesting accurate expectations may not always offer advantages. Regardless, individuals' expectations of inclusion, and the subsequent effects on relational evaluation, should be considered in future theorizing and research on the effects of social exclusion.

AUTHOR CONTRIBUTIONS

EW, JW, and MB each contributed to the theoretical arguments in this article.

REFERENCES

Arriaga, X. B., Capezza, N. M., Reed, J. T., Wesselmann, E. D., and Williams, K. D. (2014). With partners like you, who needs strangers? Ostracism involving a romantic partner. *Pers. Relat.* 21, 557–569. doi: 10.1111/pere.12048

Ayduk, Ö., Gyurak, A., and Luerssen, A. (2008). Individual differences in the rejection– aggression link in the hot sauce paradigm: the case of rejection sensitivity. *J. Exp. Soc. Psychol.* 44, 775–782. doi: 10.1016/j.jesp.2007.07.004

Baumeister, R. F., and Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* 117, 497–529. doi: 10.1037/0033-2909.117.3.497

Baumeister, R. F., Twenge, J. M., and Nuss, C. K. (2002). Effects of social exclusion on cognitive processes: anticipated aloneness reduces intelligent thought. *J. Pers. Soc. Psychol.* 83, 817–827. doi: 10.1037/0022-3514.83.4.817

Bernstein, M. J., and Claypool, H. M. (2012). Social exclusion and pain sensitivity why exclusion sometimes hurts and sometimes numbs. *Pers. Soc. Psychol. Bull.* 38, 185–196. doi: 10.1177/0146167211422449

Bernstein, M. J., Claypool, H. M., Young, S. G., Tuscherer, T., Sacco, D. F., and Brown, C. M. (2013). Never let them see you cry self-presentation as a moderator of the relationship between exclusion and self-esteem. *Pers. Soc. Psychol. Bull.* 39, 1293–1305. doi: 10.1177/0146167213495281

Bernstein, M. J., Sacco, D. F., Brown, C. M., Young, S. G., and Claypool, H. M. (2010a). A preference for genuine smiles following social exclusion. *J. Exp. Soc. Psychol.* 46, 196–199. doi: 10.1016/j.jesp.2009.08.010

Bernstein, M. J., Sacco, D. F., Young, S. G., Hugenberg, K., and Cook, E. (2010b). Being "in" with the in-crowd: the effects of social exclusion and inclusion are enhanced by the perceived essentialism of ingroups and outgroups. *Pers. Soc. Psychol. Bull.* 36, 999–1009. doi: 10.1177/0146167210376059

Bernstein, M. J., Young, S. G., Brown, C. M., Sacco, D. F., and Claypool, H. (2008). Adaptive responses to social exclusion: social rejection improves detection of real and fake smiles. *Psychol. Sci.* 19, 981–983. doi: 10.1111/j.1467-9280.2008.02187.x

Böckler, A., Hömke, P., and Sebanz, N. (2014). Invisible man: exclusion from shared attention affects gaze behavior and self-reports. *Soc. Psychol. Personal. Sci.* 5, 140–148. doi: 10.1177/1948550613488951

Buckley, K. E., Winkel, R. E., and Leary, M. R. (2004). Reactions to acceptance and rejection: effects of level and sequence of relational evaluation. *J. Exp. Soc. Psychol.* 40, 14–28. doi: 10.1016/S0022-1031(03)00064-7

Bushman, B. J., and Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: does self-love or self-hate lead to violence? *J. Pers. Soc. Psychol.* 75, 219–229. doi: 10.1037/0022-3514.75.1.219

Cursan, A., Bernstein, M. J., Pascual, A., and Félonneau, M. L. (2016). Impact of gendered ingroup/outgroup ostracism on women's academic performances. *J. Soc. Psychol.* doi: 10.1080/00224545.2016.1215966. [Epub ahead of print].

DeWall, C. N., Maner, J. K., and Rouby, D. A. (2009). Social exclusion and early-stage interpersonal perception: selective attention to signs of acceptance. *J. Pers. Soc. Psychol.* 96, 729–741. doi: 10.1037/a0014634

Dotan-Eliaz, O., Sommer, K. L., and Rubin, Y. S. (2009). Multilingual groups: effects of linguistic ostracism on felt rejection and anger, coworker attraction, perceived team potency, and creative performance. *Basic Appl. Soc. Psychol.* 31, 363–375. doi: 10.1080/01973530903317177

Downey, G., and Feldman, S. I. (1996). Implications of rejection sensitivity for intimate relationships. *J. Pers. Soc. Psychol.* 70, 1327–1343. doi: 10.1037/0022-3514.70.6.1327

Gardner, W. L., Pickett, C. L., and Brewer, M. B. (2000). Social exclusion and selective memory: how the need to belong influences memory for social events. *Pers. Soc. Psychol. Bull.* 26, 486–496. doi: 10.1177/0146167200266007

Geller, D. M., Goodstein, L., Silver, M., and Sternberg, W. C. (1974). On being ignored: the effects of the violation of implicit rules of social interaction. *Sociometry* 37, 541–556. doi: 10.2307/2786426

Gonsalkorale, K., and Williams, K. D. (2007). The KKK won't let me play: ostracism even by a despised outgroup hurts. *Eur. J. Soc. Psychol.* 37, 1176–1185. doi: 10.1002/ejsp.392

Goodwin, S. A., Williams, K. D., and Carter-Sowell, A. R. (2010). The psychological sting of stigma: the costs of attributing ostracism to racism. *J. Exp. Soc. Psychol.* 46, 612–618. doi: 10.1016/j.jesp.2010.02.002

Hales, A. H., Williams, K. D., and Rector, J. (in press). *Alienating the Audience: How Abbreviations Hamper Scientific Communication*. APS Observer.

Hitlan, R. T., Kelly, K. M., Schepman, S., Schneider, K. T., and Zárate, M. A. (2006). Language exclusion and the consequences of perceived ostracism in the workplace. *Group Dyn. Theor. Res. Pract.* 10:56. doi: 10.1037/1089-2699.10.1.56

Jones, E. E., Carter-Sowell, A. R., and Kelly, J. R. (2011). Participation matters: psychological and behavioral consequences of information exclusion in groups. *Group Dyn. Theor. Res. Pract.* 15, 311–325. doi: 10.1037/a0025547

Jones, E. E., Carter-Sowell, A. R., Kelly, J. R., and Williams, K. D. (2009). I'm out of the loop': Ostracism through information exclusion. *Group Process. Intergroup Relat.* 12, 157–174. doi: 10.1177/1368430208101054

Kerr, N. L., and Levine, J. M. (2008). The detection of social exclusion: evolution and beyond. *Group Dyn. Theor. Res. Pract.* 12, 39–52. doi: 10.1037/1089-2699.12.1.39

Klages, S. V., and Wirth, J. H. (2014). Excluded by laughter: laughing until it hurts someone else. *J. Soc. Psychol.* 154, 8–13. doi: 10.1080/00224545.2013.843502

Koudenburg, N. (2014). *Conversational Flow*. Doctoral dissertation, University of Groningen.

Koudenburg, N., Postmes, T., and Gordijn, E. H. (2011). Disrupting the flow: how brief silences in group conversations affect social needs. *J. Exp. Soc. Psychol.* 47, 512–515. doi: 10.1016/j.jesp.2010.12.006

Koudenburg, N., Postmes, T., and Gordijn, E. H. (2013). Conversational flow promotes solidarity. *PLoS ONE* 8:e78363. doi: 10.1371/journal.pone.0078363

Leary, M. R. (1990). Responses to social exclusion: social anxiety, jealousy, loneliness, depression, and low self-esteem. *J. Soc. Clin. Psychol.* 9, 221–229. doi: 10.1521/jscp.1990.9.2.221

Leary, M. R. (1999). Making sense of self-esteem. *Curr. Dir. Psychol. Sci.* 8, 32–35. doi: 10.1111/1467-8721.00008

Leary, M. R., and Baumeister, R. F. (2000). "The nature and function of self-esteem: sociometer theory," in *Advances in Experimental Social Psychology*, Vol. 32, ed M. P. Zanna (San Diego, CA: Academic Press), 1–62.

Leary, M. R., Cottrell, C. A., and Phillips, M. (2001). Deconfounding the effects of dominance and social acceptance on self-esteem. *J. Pers. Soc. Psychol.* 81, 898–909. doi: 10.1037/0022-3514.81.5.898

Lieberman, M. D. (2013). *Social: Why Our Brains are Wired to Connect*. New York, NY: Crown Publishers.

Maner, J. K., DeWall, C. N., Baumeister, R. F., and Schaller, M. (2007). Does social exclusion motivate interpersonal reconnection? Resolving the "porcupine problem." *J. Pers. Soc. Psychol.* 92, 42–55. doi: 10.1037/0022-3514.92.1.42

Mead, N. L., Baumeister, R. F., Stillman, T. F., Rawn, C. D., and Vohs, K. D. (2011). Social exclusion causes people to spend and consume strategically in the service of affiliation. *J. Cons. Res.* 37, 902–919. doi: 10.1086/656667

Nezlek, J. B., Wesselmann, E. D., Wheeler, L., and Williams, K. D. (2012). Ostracism in everyday life. *Group Dyn. Theor. Res. Pract.* 16, 91–104. doi: 10.1037/a0028029

Pfundmair, M., DeWall, C. N., Fries, V., Geiger, B., Krämer, T., Krug, S., et al. (2015). Sugar or spice: using I^3 metatheory to understand how and why glucose reduces rejection-related aggression. *Aggress. Behav.* 41, 537–543. doi: 10.1002/ab.21593

Pickett, C. L., and Gardner, W. L. (2005). "The social monitoring system: enhanced sensitivity to social cues as an adaptive response to social exclusion," in *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying*, eds K. D. Williams, J. P. Forgas, and W. von Hippel (New York, NY: Psychology Press), 213–226.

Pickett, C. L., Gardner, W. L., and Knowles, M. (2004). Getting a cue: the need to belong and enhanced sensitivity to social cues. *Pers. Soc. Psychol. Bull.* 30, 1095–1107. doi: 10.1177/0146167203262085

Riva, P., and Eck, J. (2016). "The many faces of social exclusion," in *Social Exclusion: Psychological Approaches to Understanding and Reducing Its Impact*, eds P. Riva and J. Eck (Springer International Publishing), 9–15.

Riva, P., Montali, L., Wirth, J. H., Curioni, S., and Williams, K. D. (2016). Chronic social exclusion and evidence for the resignation stage: an empirical investigation. *J. Soc. Pers. Relat.* doi: 10.1177/0265407516644348. [Epub ahead of print].

Riva, P., Wesselmann, E. D., Wirth, J. H., Carter-Sowell, A. R., and Williams, K. D. (2014). When pain does not heal: the common antecedents and consequences of chronic social and physical pain. *Basic Appl. Soc. Psych.* 36, 329–346. doi: 10.1080/01973533.2014.917975

Rudert, S. C., and Greifeneder, R. (2016). When it's okay that I don't play: social norms and the situated construal of social exclusion. *Pers. Soc. Psychol. Bull.* 42, 955–969. doi: 10.1177/0146167216649606

Sacco, D. F., Wirth, J. H., Hugenberg, K., Chen, Z., and Williams, K. D. (2011). The world in black and white: ostracism enhances the categorical perception of social information. *J. Exp. Soc. Psychol.* 47, 836–842. doi: 10.1016/j.jesp.2011.03.001

Shilling, A. A., and Brown, C. M. (2016). Goal-driven resource redistribution: an adaptive response to social exclusion. *Evol. Behav. Sci.* 10, 149–167. doi: 10.1037/ebs0000062

Smart Richman, L., and Leary, M. R. (2009). Reactions to discrimination, stigmatization, ostracism, and other forms of interpersonal rejection: a multimotive model. *Psychol. Rev.* 116, 365–383. doi: 10.1037/a0015250

Somerville, L. H., Heatherton, T. F., and Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nat. Neurosci.* 9, 1007–1008. doi: 10.1038/nn1728

Spoor, J., and Williams, K. D. (2007). "The evolution of an ostracism detection system," in *The Evolution of the Social Mind: Evolutionary Psychology and Social Cognition*, eds J. P. Forgas, M. Haselton, and W. von Hippel (New York, NY: Psychology Press), 279–292.

Tuscherer, T., Sacco, D. F., Wirth, J. H., Claypool, H. M., Hugenberg, K., and Wesselmann, E. D. (2016). Responses to exclusion are moderated by its perceived fairness. *Eur. J. Soc. Psychol.* 46, 280–293. doi: 10.1002/ejsp.2152

Twenge, J. M., Baumeister, R. F., Tice, D. M., and Stucke, T. S. (2001). If you can't join them, beat them: effects of social exclusion on aggressive behavior. *J. Pers. Soc. Psychol.* 81, 1058–1069. doi: 10.1037/0022-3514.81.6.1058

Twenge, J. M., and Campbell, W. K. (2003). "Isn't it fun to get the respect that we're going to deserve?" Narcissism, social rejection, and aggression. *Pers. Soc. Psychol. Bull.* 29, 261–272. doi: 10.1177/0146167202239051

Wesselmann, E. D., Butler, F. A., Williams, K. D., and Pickett, C. L. (2010). Adding injury to insult: unexpected rejection leads to more aggressive responses. *Aggress. Behav.* 36, 232–237. doi: 10.1002/ab.20347

Wesselmann, E. D., Cardoso, F. D., Slater, S., and Williams, K. D. (2012a). To be looked at as though air: civil attention matters. *Psychol. Sci.* 23, 166–168. doi: 10.1177/0956797611427921

Wesselmann, E. D., Grzybowski, M. R., Steakley-Freeman, D. M., DeSouza, E. R., Nezlek, J. B., and Williams, K. D. (2016). "Social exclusion in everyday life," in *Social Exclusion: Psychological Approaches to Understanding and Reducing Its Impact*, eds P. Riva and J. Eck (Springer International Publishing), 3–23.

Wesselmann, E. D., Nairne, J. S., and Williams, K. D. (2012b). An evolutionary social psychological approach to studying the effects of ostracism. *J. Soc. Evol. Cult. Psychol.* 6, 309–328. doi: 10.1037/h0099249

Wesselmann, E. D., Ren, D., and Williams, K. D. (2015). Motivations for responses to ostracism. *Front. Psychol.* 6:40. doi: 10.3389/fpsyg.2015.00040

Wesselmann, E. D., Wirth, J. H., Pryor, J. B., Reeder, G. D., and Williams, K. D. (2013). When do we ostracize? *Soc. Psychol. Personal. Sci.* 4, 108–115. doi: 10.1177/1948550612443386

Williams, K. D. (2009). "Ostracism: effects of being excluded and ignored," in *Advances in Experimental Social Psychology*, ed M. Zanna (New York, NY: Academic Press), 275–314.

Williams, K. D., Cheung, C. K., and Choi, W. (2000). Cyberostracism: effects of being ignored over the Internet. *J. Pers. Soc. Psychol.* 79, 748–762. doi: 10.1037/0022-3514.79.5.748

Williams, K. D., and Wesselmann, E. D. (2011). "The link between ostracism and aggression," in *The Psychology of Social Conflict and Aggression*, eds J. P. Forgas, A. W. Kruglanski, and K. D. Williams (New York, NY: Psychology Press), 37–51.

Wirth, J. H., Bernstein, M. J., Wesselmann, E. D., and LeRoy, A. S. (2017). Social cues establish expectation of rejection and affect the response to being rejected. *Group Process. Intergroup Relat.* 20, 32–51. doi: 10.1177/1368430215596073

Wirth, J. H., Sacco, D. F., Hugenberg, K., and Williams, K. D. (2010). Eye gaze as relational evaluation: averted eye gaze leads to feelings of ostracism and relational devaluation. *Pers. Soc. Psychol. Bull.* 36, 869–882. doi: 10.1177/0146167210370032

Wirth, J. H., and Williams, K. D. (2009). "They don't like our kind": consequences of being ostracized while possessing a group membership. *Group Process. Intergroup Relat.* 12, 111–127. doi: 10.1177/1368430208098780

Conflict of Interest Statement: 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.

Copyright © 2017 Wesselmann, Wirth and Bernstein. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 December 2016
doi: 10.3389/fpsyg.2016.01909

Expectations and Decisions in the Volunteer's Dilemma: Effects of Social Distance and Social Projection

Joachim I. Krueger[1], Johannes Ullrich[2]* and Leonard J. Chen[3]*

[1] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA, [2] Department of Psychology, University of Zurich, Zurich, Switzerland, [3] Public Service Division, Singapore, Singapore

In a *Volunteer's Dilemma* (VoD) one individual needs to bear a cost so that a public good can be provided. Expectations regarding what others will do play a critical role because they would ideally be negatively correlated with own decisions; yet, a social-projection heuristic generates positive correlations. In a series of 2-person-dilemma studies with over 1,000 participants, we find that expectations are indeed correlated with own choice, and that people tend to volunteer more than game-theoretic benchmarks and their own expectations would allow. We also find strong evidence for a social-distance heuristic, according to which a person's own probability to volunteer and the expectation that others will volunteer decrease as others become socially more remote. Experimentally induced expectations make opposite behavior more likely, but respondents underweight these expectations. As a result, there is a small but systematic effect of over-volunteering among psychologically close individuals.

Keywords: social dilemma, prosociality, expectation, rationality

OPEN ACCESS

Edited by:
Mario Gollwitzer,
University of Marburg, Germany

Reviewed by:
Robert Gaschler,
FernUniversität in Hagen, Germany
Erik Willem De Kwaadsteniet,
Leiden University, Netherlands
Dorothee Mischkowski,
University of Göttingen, Germany

****Correspondence:***
Joachim I. Krueger
joachim@brown.edu
Johannes Ullrich
j.ullrich@psychologie.uzh.ch

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *29 August 2016*
Accepted: *22 November 2016*
Published: *06 December 2016*

Citation:
Krueger JI, Ullrich J and Chen LJ (2016) Expectations and Decisions in the Volunteer's Dilemma: Effects of Social Distance and Social Projection. Front. Psychol. 7:1909. doi: 10.3389/fpsyg.2016.01909

INTRODUCTION

> "That love as such may be unable to settle a conflict can be shown by considering a harmless test case, which may pass as representative of more serious ones. Tom likes the theater and Dick likes dancing. Tom lovingly insists on going to a dance while Dick wants for Tom's sake to go to the theater. This conflict cannot be settled by love; rather, the greater the love, the stronger will be the conflict. There are only two solutions; one is the use of emotion, and ultimately of violence, and the other is the use of reason, of impartiality, of reasonable compromise."
>
> Sir Karl Popper (1945/2011, p. 441)

Surviving and flourishing in the natural and the cultural world requires decision-making skills. In games against nature, humans and other animals seek to do whatever ensures the survival of their physical selves and the genes they carry (Buss, 1999). They need to forage efficiently in environments characterized by uncertainty, scarcity, and an indifference to their welfare. In social games, which often involve self-interested and only sometimes empathic conspecifics, humans need to predict what these others will do when they know that these others are also trying to figure out what they themselves will do (Hoffrage and Hertwig, 2012). Social games demand the kind of strategic reasoning that generates and makes use of expectations in a dynamical way. These games demand – as Popper realized – reason, impartiality, and compromise.

What sort of reason is it? Game theory offers a formal paradigm for the description of social games or dilemmas and for derivations of rational choice (Von Neumann and Morgenstern, 1947; Luce and Raiffa, 1957; Binmore, 2007). Orthodox game theory does not face the problem of expectation squarely; it finesses the problem of other minds by defining it away. Consider

game-theory's iconic game, the prisoner's dilemma, or PD. The person (or 'agent' or 'player') who is rational in the game-theoretic sense defects, hoping perhaps – though not expecting – that others will cooperate. This player recognizes defection as the dominating strategy. Whatever the other player (in a 2-person game) does, this player fares better defecting. Unilateral defection pays more (or penalizes less) than bilateral cooperation, and bilateral defection pays more than unilateral cooperation. To find the rational response, the player only needs to subtract one payoff from another, do this twice, and note that the ordinal result is the same. In other words, the player only needs to understand that defection is the "sure thing" (Tversky and Shafir, 1992). As the direction of the difference is the same regardless of the expected probability of the other player cooperating (or defecting), the concept of expectation drops out.

Noting the psychological barrenness of classic game theory and worrying about its limited descriptive success (i.e., the finding that many reasonable people cooperate in the PD), revisionist theorists have reintroduced expectations as a necessary determinant of rational choice (Pruitt and Kimmel, 1977; Monterosso and Ainslee, 2003; Rapoport, 2003). Research has shown that many individuals cooperate on the condition that there is evidence or a good expectation that the other person will also cooperate (Gintis, 2000; Fischbacher et al., 2001; Nielsen et al., 2014).

A related line of research suggests that many individuals expect others to choose the same strategy that they themselves will choose, and that they therefore end up choosing cooperation (Fischer, 2009; Krueger, 2013, 2014). According to this alternative perspective on social dilemmas, the generation of behavioral expectations and their effects on own choice is neither unnecessary nor irrational. Since the days of Pascal (1995/1669) and Bernoulli (1954/1738), the multiplicative integration of expectations and values (i.e., payoffs) lies at the heart of most theories of rational choice (e.g., Ajzen and Fishbein, 2008). These theories assume that people are either able to multiply and that they choose well, or that at least their choices fit the predictions made from explicit multiplications of expectations and values, that is, people act at least *as if* they were making the requisite calculations (Berg and Gigerenzer, 2010).

The research we report in this article is concerned with the volunteer's dilemma, or VoD, which belongs to a class of games in which rational agents would wish to choose opposite strategies. These dilemmas are known as anti-coordination games. Here each player's goal is to mismatch the other player's strategy, which raises particular psychological challenges (Abele et al., 2014). As in other social dilemmas (including the PD), there is a choice between one strategy that favors the self and another strategy that favors the other person or the group (Archetti and Scheuring, 2011). The outcome depends both on one's own choice and the choice of the other, and there is an inequality: the individual and the collective outcome of mutual cooperation are better than the outcome of mutual defection (Dawes, 1980; Krueger et al., 2016). Yet, there is an incentive to defect, which raises the specter of the destructive outcome of mutual defection (Hardin, 1968). Whereas the structure of the PD makes it easy for the game-theoretic rationalist to understand that defection dominates cooperation, the VoD offers no dominating strategy. This feature is a definitional property of games that yield best results when the two agents choose different strategies, such as the game of chicken (Rapoport and Chammah, 1966, which is also know as the hawk-dove game, or its multi-player extension, the crowding game; Alpern and Reyniers, 2001). Game theory responds to this challenge with the concept of the mixed-strategy Nash equilibrium, which is designed to withhold from the other person any incentive to change strategy. Again, expectations are unnecessary for the derivation of the Nash equilibrium strategy.

Consider the structure of the VoD as displayed in **Figure 1**. Volunteering yields the outcome (or payoff) "R," which stands for "Reward" (after Rapoport, 1967). R is obtained regardless of the other player's choice. Defection yields payoff "T" (for "Temptation") if the other player volunteers, but payoff "P" ("Penalty") if the other defects. There is a social dilemma because $T > R > P$. Situations satisfying the definition of the VoD crop up throughout social life whenever a division of labor and responsibility is not regulated by contract or custom. Lecturers, for example, hope for a student to volunteer to speak in class and thereby ignite discussion; victims of emergency hope that one person will help; soldiers on the battlefield sometimes need one comrade who will accept the riskiest mission so that the others may live.

When communication and coordination are impossible, each individual must decide independently what to do. Diekmann (1985) derived the mixed-strategy Nash equilibrium probability of volunteering as $(R - P)/(T - P)$. The difference $R - P$ can be thought of as the psychic benefit of volunteering, but also as the potential cost of not volunteering. The difference $T - P$ represents the total cost of mutual defection, which is the sum of $T - R$ (i.e., the temptation to defect) and $R - P$. We consider it psychologically implausible that people approach

Player 1 , Player 2	Option A	Option B
Option A	R,R 1, 1	R,T 1, 2
Option B	T,R 2, 1	P,P 0, 0

FIGURE 1 | Payoff Matrix of the Volunteer's Dilemma. Option A is to volunteer; Option B is to abstain.

a social dilemma without wondering what other individuals will do. Even an orthodox game theorist assumes (or expects) that a Nash-playing person will assume that other individuals will do likewise. This is a non-trivial expectation because even though deviating from Nash cannot improve one's own payoffs, it can hurt the payoffs of the other (Krueger et al., 2016, unpublished). In short, the game-theoretic approach postulates the belief in common knowledge, which is tantamount to a multi-level shared expectation (Thomas et al., 2014). Game theory assumes that players are not motivated by malice and that they do not expect others to be so motivated.

Expectations: The Social Projection Hypothesis

The questions of whether people form expectations about others in social dilemmas and whether such expectations affect strategic decisions are separable. With regard to the first question, there is empirical support for the idea that people form expectations projectively: they think that others are likely to choose whichever strategy they themselves prefer. Dawes et al. (1977) presented evidence for this hypothesis (see also Messé and Sivacek, 1979) and Dawes (1989) derived a Bayesian rationale for why people *should* use their own strategic choice as a projective cue to predict the choices of others, and proved by backward induction that even a sample of one ought not be ignored lest a sample of any size would have to be ignored. This logic is particularly compelling in an information-poor environment such as an anonymous one-shot social dilemma.

With regard to the second question, it has been argued that once projection is admitted as a judgment heuristic, it cannot be ignored as a decision heuristic (Krueger and Acevedo, 2005). In the PD, for example, the rational expectation that most others – by definition – are more likely to make the same instead of a different choice will leave a person caught between the prospects of mutual cooperation and mutual defection. Being able to only predict mutuality by using the projection heuristic, a self-interested player has no reason *not* to choose cooperation. Choosing cooperation does not imply a magical belief that the other person's behavior can be influenced but simply reflects respect for the statistical rule that one's own choice is diagnostic of the choices of most others (Krueger, 2013; Krueger and Acevedo, 2005; Krueger et al., 2012). Social projection is beneficial in the PD because mutual cooperation is best for both the individual and the group, whereas in the VoD, projection is problematic because mutual cooperation (2R) is worse than unilateral cooperation (T+R). Ideally, a player would choose whichever strategy the other player is not choosing. If Tom knows that Dick volunteers, Tom defects. If Tom knows that Dick defects, Tom volunteers. The structure of the VoD thus challenges the human tendency to project. A player who volunteers and then estimates that the other player will also volunteer will be dissatisfied with the prospect of mutual, that is, inefficient, volunteering. A player who defects and then estimates that the other player will also defect will be unhappy with the prospect of mutual loss. In other words, these players find themselves in Popper's dilemma of love.

If the VoD does not reward social projection, one might think that projection is low or even reversed in this dilemma. Our working hypothesis, however, is that projection will be strong nonetheless. We draw this hypothesis from past research, which has shown that projection is a reliable social heuristic even under conditions discouraging its use (Krueger and Clement, 1994; Krueger, 2003). We predict that in the VoD players' strategy choices will be positively correlated with the choices expected of others.[1]

Evolution: The Social Distance Hypothesis

Classic game theory is not concerned with individual differences, identity, or social categories. The theory does not simply happen to ignore such variables. Its axioms affirm their irrelevance. There is only one standard of rational choice, and everyone is assumed to meet it. In contrast, social psychology and evolutionary psychology recognize the relevance of prosocial motives and how these motives are differentially activated by the nature of the relationships between or among actors (Murphy and Ackermann, 2014; Kurzban et al., 2015). The broadest generalization emerging from theory and data is that the probability of prosocial choice decreases with social (or psychological or genetic) distance. Hamilton's (1964) theory of inclusive fitness provides an elegant Darwinian rationale. Assuming that the survival of genes is the ultimate adaptive coin, organisms will make sacrifices if and only if the net effect on the survival of their genes is positive. Prosocial behavior will therefore decrease as the beneficiaries of these sacrifices become biologically more distant. In a classic study, Burnstein et al. (1994) showed that people come to the aid of close over distant kin in hypothetical life-and-death scenarios, whereas less serious contexts activate social norms concerning need and deservingness. Genetic relatedness is difficult to display and assess, and humans and other animals have evolved a range of cues to honestly or deceptively signal relatedness (Dawkins, 1976). Perhaps the crudest way to differentiate between close and distant others is to categorize them into ingroups and outgroups. The general finding is that people like their ingroups more than outgroups (Krueger and DiDonato, 2008), describe them in more favorable terms, and – importantly – are more willing to help ingroup than outgroup members in need (Rabbie and Horwitz, 1969; Tajfel and Turner, 1979; DiDonato et al., 2011).

From the perspective of biology, anthropology, and psychology, "bounded prosociality" is a stylized fact (De Dreu et al., 2015). Its robustness presents a challenge to traditional game theory. There is much evidence to show that people cooperate more readily with presumed ingroup members than outgroup members in a variety of social dilemmas (Balliet et al., 2014). Importantly, the increased willingness to cooperate in the context of "parochial morality" comes with the expectation that ingroup members, but not outgroup members, will also cooperate (Yamagishi and Kiyonari, 2000; Brewer, 2008). In other words, differential projection (Robbins and Krueger, 2005)

[1]If there is support for the projection hypothesis in the VoD, we will have an argument against the idea that people project strongly in the PD only *in order to rationalize their own cooperative desires.*

tends to be accurate. Extrapolating from this research, we hypothesize that people's readiness to volunteer and their expectations that others will volunteer both diminish over social distance. Although such a decline runs counter to the precepts of traditional game theory, it is consistent with certain social preference models of interdependent behavior (e.g., Van Lange, 1999; Fehr et al., 2005; Archetti, 2009).

Archetti (2009) developed a social-preference model to quantitatively predict the probability of volunteering for degrees of social distance. With our payoff notation, Archetti's (2009, p. 476) equation becomes $p_v = 1 - \frac{T-R}{(T-P)\cdot[1+(1-d)]}$. The probability of volunteering, p_v, increases as the temptation to defect, T − R, or the cost of mutual defection, T − P, decrease and as social distance, d, increases. The parameter d captures the idea that the utility of volunteering is high to the extent that the other person is socially or genetically close to the self. Consider the payoffs in **Figure 1**, namely T = 2, R = 1, and P = 0. For maximum distance (d = 1), we find that $p_v = 0.5$, which is the conventional Nash equilibrium. Neither orthodox game theory nor a biologically informed social-preference theory would assume a probability of volunteering below this benchmark.[2] For zero distance $p_v = 0.75$. Here, the player weights the outcomes of the other as much as his or her own outcomes, and if both players do this, the sum of their outcomes is maximized. Note, however, that this is not an equilibrium in the Nash sense. A player who knows or expects the other to volunteer with a high p_v might choose to defect for sure and thereby increase his or her payoff and reduce the other's. In other words, using this 'superrational' strategy (Diekmann, 1985) requires the expectation that the other player will do the same.

A Costly Error: The Over-Volunteering Hypothesis

Our third hypothesis is more subtle and thus riskier. We predict that many individuals will volunteer too much relative to formal standards and relative to the implications of their own expectations regarding others' choices. They will, in other words, stumble into Popper's dilemma of love. How might this happen? We submit that the social-distance heuristic is frugal in the sense that it has no non-monotonic provisos (Gigerenzer and Gaissmaier, 2011). There is no check as to whether there may be too much volunteering. Not having such a proviso works well in social dilemmas where mutual cooperation is the most efficient collective strategy (i.e., were 2R > T + P [or T + S]). In the VoD, however, heuristically thinking individuals may choose to volunteer for a very close other without working out the implications. As both individuals have this tendency, the outcome is inefficient. In other words, we predict that Archetti's social preference model will offer a good description of volunteering over social distance, but that against this background of adaptiveness, there will be a systematic error precisely where individuals would want to avoid it the most.

[2]This prediction refers to group averages. It is conceivable to find strict defectors whose primary goal is to exploit the other, or, equivalently, to ensure not to earn a lower payoff than the other.

When adding expectations to the picture, the possibility of over-volunteering becomes more poignant. If, as we hypothesize, people will be most likely to volunteer when the other is psychologically close, and if, as we also hypothesize, people project their own choices most strongly onto those who are close, then we will find that respondents over-volunteer even by the lights of their own expectations of reciprocity. To illustrate this hypothesis, imagine a pair of siblings. Both want to 'do the right thing' and sacrifice for the other. At the same time, they predict that their sibling is equally willing to make that sacrifice. Yet, they choose to volunteer. This outcome, if obtained, would suggest that projective predictions are difficult to alter. The player cannot escape the dilemma by defecting because this would suggest the worst personal and collective outcome. To avoid over-volunteering, the person would have to find a way to predict that the other person is less likely to volunteer than the self. This, in turn, might be a difficult psychological maneuver because it would suggest that the self is a more socially responsible person than the other. In doing so, it would undermine the perception of social closeness (there is, however, evidence for such self-enhancement in volunteering, Heck and Krueger, 2016, unpublished).

Research Overview

We tested these hypotheses in three studies. In study 1, we sought to demonstrate the social-distance effect and provide evidence for over-volunteering at very short social distances, as evaluated against a game-theoretic standard. In study 2, we considered a full range of social distances and introduced respondents' expectations. Here, we tested all three hypotheses (social projection, social distance, and over-volunteering) over multiple samples. In study 3, we manipulated expectations experimentally. Assuming that expectations are not epiphenomenal to behavior, we predicted that respondents would consult expectations when making a decision, but that the effect would be limited and result in over-volunteering.

STUDY 1: SOCIAL DISTANCE AND OVER-VOLUNTEERING

Undergraduate mostly female students ($N = 250$) in a 1st-year lecture course on social psychology at a German-language university in Switzerland took part in a classroom experiment. No demographic data were collected. Students received instructions over the microphone and were shown the following information on a large screen. Instructions read that "the goal of this experiment is to illustrate, with the help of your imagination, a social dilemma, that is a game for at least two persons, in which the consequences depend on the decisions of all participants. You will be asked to make a hypothetical decision that may entail that you or someone else will hypothetically receive an electric shock. Participation is anonymous and voluntary." Next, participants were asked to imagine gradations of social distance using a method developed by Jones and Rachlin (2006) which asks participants to create a mental ranking of 100 people with rank #1 corresponding to a close friend or relative and rank

#100 corresponding to a superficial acquaintance (see below for a detailed description of this method in the context of Study 2). Then the payoff matrix of the VoD was shown and explained. Students learned that they would receive 1 (hypothetical) electric shock if they volunteered, receive no shock if they did not volunteer while the other person did, and receive 2 shocks if neither they nor the other person volunteered. That is, the payoffs were $T = 0$, $R = -1$, and $P = -2$. This payoff structure is a simple linear transformation of the canonical structure discussed earlier and displayed in **Figure 1**. Using an online response interface, all participants made two binary decisions to either select Option A or Option B, which, respectively, amounted to volunteering and defecting. They made the first decision under the presumption that they were paired with the person of the lowest social distance (person #1 on the ranked list), and they made the second decision under the presumption of being paired with the person of the greatest social distance (person #100 on the list).

The results supported the social distance and the over-volunteering hypotheses. For the closest distance (rank #1), 87% volunteered. The 95% confidence interval, CI: [82; 91] excluded the equilibrium value of 75%, which would maximize joint outcomes. For the greatest distance (rank #100), 68% volunteered, and the 95% CI [62; 73] excluded its corresponding Nash equilibrium value of 50%, that is, the strategy of the rational, self-interested individual.

This was first evidence for the social distance hypothesis. Moreover, when compared with game-theoretic benchmarks, there was evidence for over-volunteering not only for a VoD involving close others but also involving distant others. Expectations were neither measured nor manipulated and no intermediate levels of social distance were considered. We designed a multi-sample study to address these issues.

STUDY 2: A CONTINUUM OF SOCIAL DISTANCE AND EXPECTATIONS

The goal of this study was to test the social distance, social projection, and over-volunteering hypotheses in the context of social expectations. We wanted to see whether people over-volunteer (at close distance) even in light of their own expecations regarding the other's decision to volunteer. As discussed earlier, this prediction followed from the social projection hypothesis. In addition to tests of these three main hypotheses, the data also allowed us to ask whether respondents tended to think that they themselves were more likely to volunteer than others, and whether such a tendency might be moderated by social distance. If obtained, such a self-enhancement effect ("I volunteer more than the other"; Heck and Krueger, 2015) would constrain over-volunteering in the sense that it would make it less likely that people would volunteer with a high probability *and* expect the same from the other.

Methods

Participants

We recruited a total of 703 participants in five samples, two of which came from a university campus in the Northeastern United States. Sample 1 was collected in the spring of 2013 and included 80 women and 80 men with a median age of 20 years. Sample 2 was collected in the spring of 2014 and included 94 women and 114 men with a median age of 20 years. Sample 3 was collected in the summer of 2014 at a campus in the German-speaking part of Switzerland. This sample included 62 men and 56 women (median age = 24). Samples 4 and 5 were collected in the fall of 2014 during a lecture class at the same Swiss University. Sample 4 (79 women and 26 men, median age = 21) received a dilemma with positive payoffs, whereas Sample 5 (76 women and 32 men, median age = 21) worked with negative payoffs (see below). Assignment to Samples 4 or 5 was random. All five samples shared nearly identical experimental procedures, which allowed us to analyze the data using a single statistical model in which the sample was entered as a potential moderator variable. This method offered an internal test of replicability and provided substantial statistical power (Schimmack, 2012). We describe the procedure for the largest sample (i.e., Sample 2) and note where the others differ.

Procedure

Participants were approached on an urban college campus in the Northeastern United States. All agreed to complete a brief survey on interdependent behavior. Each of 26 surveyors recruited eight respondents. The recruiters were enrolled in a laboratory course on social cognition, and they explained to the respondents that the data were being collected for a class project with the possibility of publication. Recruiters ensured that each respondent was surveyed individually and in a quiet location. The recruiter provided a sheet with instructions and the survey itself in a printed packet. The surveyor stayed on site, responded to questions of clarification, and thanked and debriefed the respondents upon completion of the survey.

The procedure for Sample 3 was slightly different in that only two surveyors recruited participants and no gender quota was used. For Sample 1, there were 20 surveyors. Samples 4 and 5 were collected during a lecture class with five teaching assistants distributing the questionnaires. Participants were promised a presentation on the results in return for their voluntary participation.

Materials

Instructions stated that the survey was designed "to tap into students' intuitions regarding how they would behave in a situation in which they are interdependent with someone else. That is to say, what course of action would you choose if the outcome does not only depend on your choice but also someone else's."

The survey had three pages. On the first page, the VoD was described in neutral terms. Respondents were asked to "consider an interpersonal setting that is currently popular in studies on behavioral economics. The situation involves two individuals. Think of yourself as Person 1 and the other person as Person 2. Person 2 is anonymous with the exception of one bit of information, as you will see shortly. Both individuals must select a response at the same time and without knowledge of the other's choice." Next, the consequences of choosing Option A

and Option B – by the respondent and the other person – were described. Mutual selection of Option A would result in 1 painful electric shock for each person and mutual selection of Option B would result in two painful shocks for each person. If one person selected Option A, while the other person selected Option B, the former would receive 1 shock, while the latter would receive none. This array of payoffs reflects the canonical volunteer's dilemma; Option A amounts to volunteering, Option B to abstaining (see **Figure 1** for a normal form representation of the game and positive payoffs).

Next, the scale for social distance was introduced. Respondents read a modified version of Jones and Rachlin's (2006) scale for the measurement of social distance. They were asked "to imagine that you have made a list of the 100 people closest to you in the world ranging from your dearest friend or relative at position #1 to a mere acquaintance at #100. The person at number one would be someone you know well and is your closest friend or relative. The person at #100 might be someone you recognize and encounter but perhaps you may not even know their name. You do not have to physically create the list—just imagine that you have done so." Given this mental scale, respondents were asked to "consider five individuals from this hypothetical list (numbers 1, 25, 50, 75, and 100), and we will ask for two judgments in each case. Please note that we consider social distance to be symmetrical. However close or distant the other is to you, so you are to the other."

The second page began with instructions of how to make probability judgments. To facilitate comprehension, the vivid language of frequencies was used. "In situations like the one we consider here, people might use different strategies. Suppose the game were played a 100 times; a person might decide to select Option A a certain number of times and Option B the rest of the times. This number, X out of 100, can represent the probability with which the person chooses Option A in a given individual situation."

Roughly half of the respondents were first asked to provide judgments of the likelihood of their own choosing Option A, whereas the other half were first asked to judge the likelihood that the other person would choose Option A. Within each of these two counterbalanced conditions, roughly half of the respondents made ratings progressing from high to low social distance, whereas the remainder progressed in the opposite direction. These procedural variations did not have any effects on the response variables, nor did they moderate the effects of social distance. Thus, they were not further considered in Samples 4 and 5, in which we asked for the likelihood of their own choosing Option A first and used a low to high order for social distance.

The materials for Samples 3, 4, and 5 were exact translations of the materials for Sample 2. The main differences between materials for Sample 1 and Sample 2 were that (a) the cooperative response option was labeled "Volunteer" and the other option was labeled "Abstain" for Sample 1, whereas the neutral labels "Option A" and "Option B" were used for Sample 2, and (b) the instructions for the probability judgment were more ambiguous for Sample 1 in that participants were asked "How certain are you that you would volunteer (vs. abstain)? Write in a percentage value between 0 and 100." A final difference was that the scenario described in Sample 4 was not about an electric shock, but about pleasant electrical stimulation. For example, participants were told that if they chose Option A and the other player chose Option B, they would receive one pleasant electrical stimulation and the other player would receive two pleasant electrical stimulations.

To check comprehension, we asked participants in Sample 3 at the very end to go back to the probability of volunteering they had stated for a randomly selected level of social distance, and indicate the most likely outcome of a single game based on their probability of volunteering and their expected probability of the other player volunteering. Five options were given, namely the four outcomes defined by the payoff matrix and all outcomes equally probable. Due to an oversight we did not include the case in which two of the outcomes would be most probable (which would arise if either own probability of volunteering or expectation was equal to 0.5). This led to ambiguities for 9 out of 117 participants (8%) who correctly selected one of the two most probable outcomes. By treating these participants separately, we estimate the level of comprehension conservatively. The results reassured us that participants generally understood the game. Correct answers were given by 73 participants (62%); 33 participants (28%) gave wrong answers, and 2 participants did not answer the question.

Results

Analyses

Preliminary analyses revealed homogeneous results with the exception of Sample 4, where outcomes were framed as gains. We continue with analyses of the negative-frame VoD and return to the findings from Sample 4 later. **Figure 2** displays the distributions of volunteering as bean plots, with their widths reflecting the density of responses (Kampstra, 2008) at specific levels of social distance. To account for the skew in the data, we estimated standard errors and confidence intervals by bootstrapping. We modeled heterogeneity in the average levels of the response variables and the effects of social distance as random effects, using linear mixed models algorithms provided by the package lme4 (Bates et al., 2014) for the software R (R Core Team, 2014). To obtain standardized effect sizes, we used a function provided by LaHuis et al. (2014) which calculates the approximate explained variance at Level 1.

The Probability of Volunteering

The means shown in **Figure 2** (circles) support the social distance hypothesis. Volunteering (choosing 'Option A') became less likely as social distance increased. To model this trend, we regressed the stated probability of volunteering on social distance (coded from 1 = lowest distance, to 5 = highest distance). To account for differences between samples, we used unweighted effects coding with three indicator variables and their interactions with the social distance variable. The intercepts and the effect of social distance represent the unweighted mean intercept and slope, respectively, for the whole dataset (i.e., all samples except for Sample 4, see below).

The intercept of the regression was $b = 89.47$ and the slope was $b = -7.83$, with a 95% confidence interval (CI) [−8.92, −6.69]. With each stepwise increase in social distance, the reported

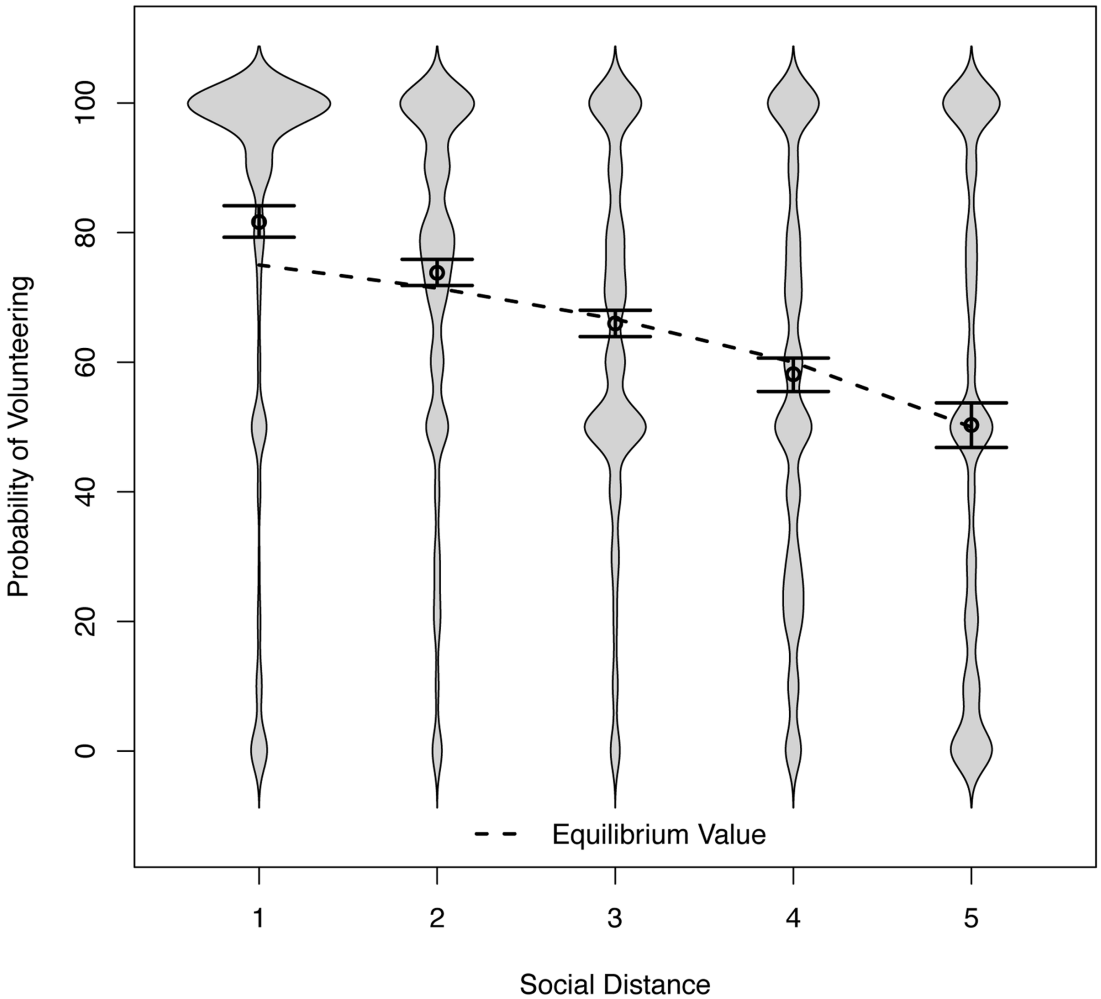


FIGURE 2 | Distributions of the probability of volunteering across social distance conditions in Study 2 (and 95% confidence intervals for the means). The lower limit of the confidence interval for the second lowest social distance condition (71.84) excludes the equilibrium value (71.43).

probability of volunteering decreased by 7.83 percentage points. The approximate explained variance at Level 1 was $R^2 = 24\%$. The individual sample intercepts and slopes from the different samples were not significantly different from the overall intercept or slope (all $|t|$ s < 1.49), which permits a joint analysis of the data.

Further analysis revealed that almost all respondents became less willing to volunteer as social distance increased. Only a few individuals produced curvilinear patterns or positive regression weights (such that the higher the social distance, the greater the stated probability of volunteering). We will return to this group when we examine the relationship between expectations and volunteering.

Figure 2 also shows the game-theoretic benchmarks for the probability of volunteering as a dotted line (Archetti, 2009). These theoretical values fit the empirical data well. There is, however, one noteworthy exception, and it corroborates the hypothesis of over-volunteering. At the two shortest social distances, respondents volunteered with a probability greater than the probability that would maximize joint outcomes (if used by both players). This mean-level difference underestimates the prevalence of over-volunteering because of the skew in the distribution. To understand how a randomly selected individual participant would choose, the width of the beans provides better guidance. For low social distance, the beans vividly illustrate the excess prosociality. In the lowest and second-lowest social distance conditions, 78 and 65% were over-volunteers, respectively (i.e., volunteering with a probability greater than the equilibrium value). The corresponding figures for those who volunteered with certainty were 59 and 31%.

Expectations of Other's Volunteering

We predicted that expectations regarding the other's probability of volunteering would also decrease over social distance, and would thus be correlated with one's own probability of volunteering. **Figure 3** shows that the data supported this prediction. In a regression of expectation on social distance, the intercept was $b = 87.97$ and the slope was $b = -10.42$, with a

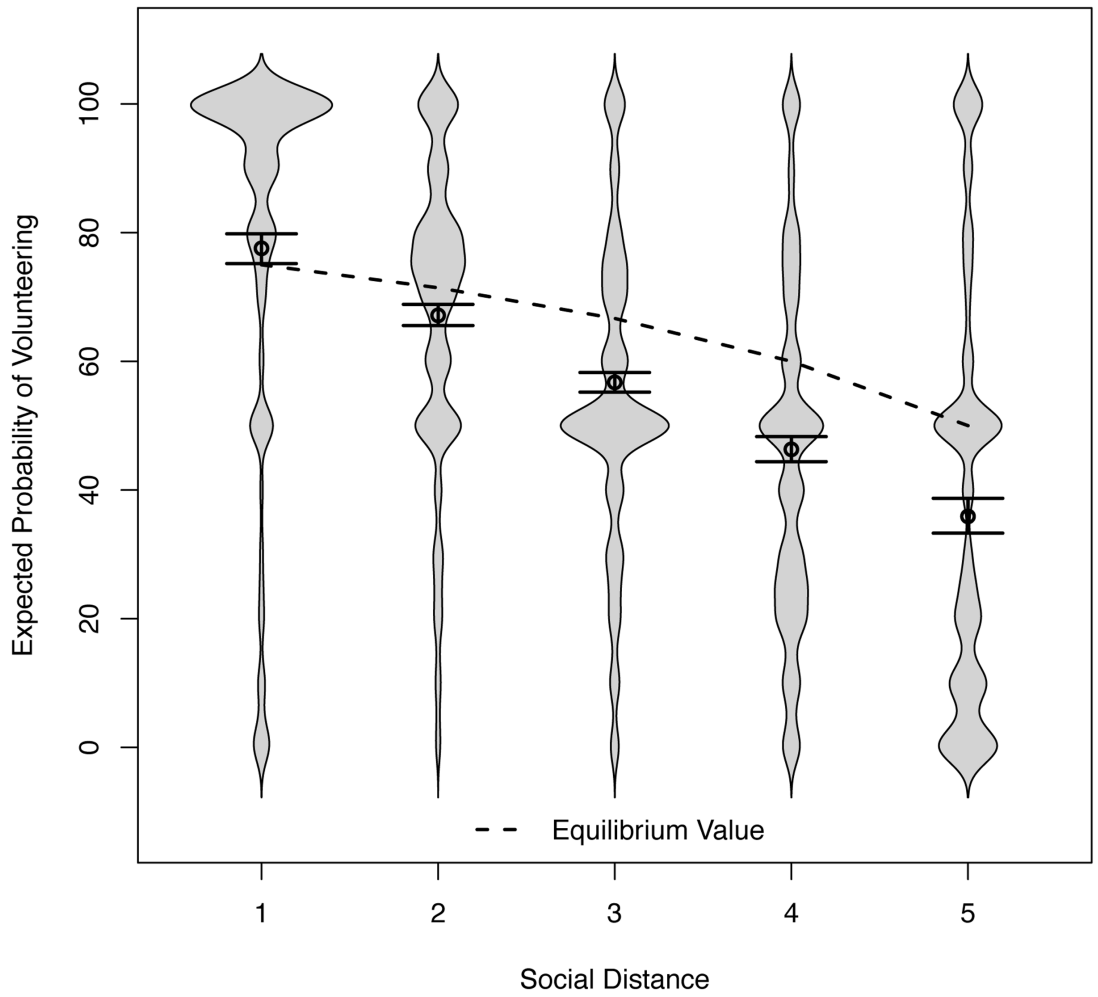


FIGURE 3 | Distributions of the expected probability of volunteering across social distance conditions in Study 2 (and 95% confidence intervals for the means). For shortest social distance, the equilibrium value (75) is below the lower limit of the confidence interval (75.19).

95% CI [−11.41, −9.32]. The approximate explained variance at Level 1 was $R^2 = 36\%$. Expected volunteering deteriorated over social distance faster than own volunteering did, thereby linking the size of the self-enhancement bias to social distance. In all but the smallest social distance conditions, respondents expected the other player to volunteer with a probability below the equilibrium. Conversely, for the closest other person, they expected others to volunteer above the equilibrium value. In other words, respondents expected the closest other player to volunteer with a greater probability than would be optimal for the dyad, mirroring the results obtained for their own volunteering. The implication is that respondents were willing to volunteer for close others with a probability that was too high in light of their own high expectations of those others volunteering.

The Relationship between Volunteering and Expectations

We tested the social projection hypothesis by regressing own volunteering on expected volunteering in a mixed model with random intercepts. As predicted, the slope of this regression was positive ($b = 0.55$, intercept = 34.95). The approximate explained variance at Level 1 was $R^2 = 34\%$. Even when considering only the data of the few participants who volunteered with a higher probability as social distance increased ($n = 85$; 14%), the slope was positive ($b = 0.20$, 95% CI [0.10, 0.29]). For these individuals, the association between behavior and expectation was weaker ($p < 0.01$) than for the majority ($b = 0.61$; 95% CI [0.58, 0.64]). The respective values of approximate explained variance at Level 1 were $R^2 = 1\%$ for the subset of participants with positive slopes and $R^2 = 44\%$ for the majority. This is strong support for the projection hypothesis. No matter which way respondents changed their willingness to volunteer over social distance, they expected others to do the same. Yet, the minority of respondents showing a positive distance effect may have had a poorer understanding of the game. In Sample 3, 73% of the participants with a negative slope for the social distance effect passed the comprehension check, whereas only 55% of participants with a positive slope did.

We assume that the correlation between own willingness to volunteer and the volunteering expected from others arises from processes of social projection rather than "introjective" mechanisms that align one's own decision with what is expected of others. It is difficult to imagine how expectations might arise without reference to one's own behavioral inclination. Indeed, if it were possible to construct such expectations early and independently, then one's own decision should be positively matched with the expected behavior of the other only when social distance is short; when individuals are paired with strangers, that is, when they act only in their own self-interest, they should do the opposite of what they expect the other to do. Yet, within each level of social distance, we find positive associations between behavior and expectation. When regressing expectations on decisions, the slope was steepest for the shortest distance ($b = 0.55$; 95% CI [0.49, 0.61]); $R^2 = 28\%$), but it was positive for the remaining four levels too (overall $b = 0.46$; 95% CI [0.42, 0.49]; $R^2 = 22\%$).

Positive Outcomes

We returned to the data obtained in Sample 4, in which payoffs were positive. Here, the slope of the regression of volunteering on social distance was flatter than it was for negative outcomes ($b = -6.40$, 95% CI [−9.12, −3.33]; $R^2 = 14\%$) and the intercept lower (79.85). As a result, the mean value of volunteering at the shortest social distance was 73.44, and the 95% CI [65.01, 81.07] included the equilibrium value (75).

We obtained similar results for the expectations regarding the probability of the other player volunteering in that the slope was flatter and the intercept lower compared with the results for negative outcomes ($b = -6.69$, 95% CI [−9.56, −3.54]; $R^2 = 14\%$; intercept $b = 77.53$). The mean value of expected probability of the other player volunteering at the shortest distance was 70.84 and the 95% CI [62.66, 78.39] included the equilibrium value (75).

For the positive outcomes too, expectations predicted volunteering ($b = 0.84$, intercept $b = 12.11$). With an approximate explained variance at Level 1 of $R^2 = 73\%$, this effect was much stronger than for negative outcomes. Within levels of social distance, own decisions predicted expectations well, and this relationship was again strongest when distance was short ($b = 0.99$ and 0.89 for the first two levels and 0.70 thereafter with respective values of explained variance $R^2 = 86$, 77, and 61%). Again, the findings suggest that participants made their own decisions to volunteer by consulting the available payoffs and weighting them by social distance, and then assuming that others would do the same.

Discussion

The results of this multi-sample study supported the main hypotheses. In support of the social-projection hypothesis, we found positive correlations between respondents' willingness to volunteer and their predictions of what the other person would do. These correlations emerged for each level of social distance, and they were strongest for short distances. It is worth noting that some "differential projection" (Robbins and Krueger, 2005), that is, a decrease of perceived similarities over social distance, is warranted because actual similarities also tend to decrease. Closely related and connected individuals share more similarities than do mere strangers. Expecting such similarities in behavior from one another is therefore a generally adaptive strategy.

As predicted, the willingness to volunteer and correspondent expectations both decreased over social distance, thereby allowing errors of over-volunteering to creep in. For the two shortest social distances, willingness to volunteer exceeded game-theoretic benchmarks. While this result suggests over-volunteering, it is not yet definitive. Respondents might rationally exceed these benchmarks if they (have reason to) believe that the others are less likely to volunteer. The clearest case for over-volunteering requires that both, own willingness to volunteer and others' expected willingness to volunteer, lie above the benchmark. We find such evidence for the shortest social distance.

Given the moral overtones of volunteering, we predicted and found evidence of self-enhancement. At each level of social distance, respondents claimed that they were, on average, more willing to volunteer than the other person. The self-enhancement bias is not a striking discovery on its own, but it is relevant in that it makes over-volunteering more difficult to detect. Had self-enhancement been any stronger, volunteers would have expected others to defect, in which case they would have expected successful (anti-)coordination to the benefit of the other.

Following theory and research on social projection, we submit that people construct expectations about others on the basis of their own behaviors rather than *vice versa* (see Van Veelen et al., 2016, for a comprehensive review of the evidence for this claim and its boundary conditions). This causal flow has good support in research on both social projection and self-enhancement (Krueger, 2007; Heck and Krueger, 2015). Yet, it is difficult to draw firm inferences in the VoD because, as in other social dilemmas, decisions and expectations are dynamically interdependent. To open a window into the potential role of expectations on volunteering decisions, we manipulated expectations in our final study. Induced expectations are available before respondents make strategic decisions (Gaschler et al., 2014). This design let us test two hypotheses: First, expectations will inversely affect volunteering decisions. Second, the effect of expectations will be smaller than full rationality demands. A consequence of this underuse of expectations is over-volunteering. Respondents will be willing to volunteer even when they expect the other person to volunteer as well.

STUDY 3: THE CAUSAL EFFECT OF EXPECTATION

We tested these hypotheses in a two-factorial repeated-measures design, in which the social distance between the respondent and the other person was either very low or very high, and in which the respondent was either led to believe that the other person was very likely or very unlikely to volunteer. Besides anticipating a replication of the social distance effect, we predicted that respondents would be more willing to defect when the other was likely to volunteer than if the other was unlikely to volunteer. In other words, we predicted an effect of expectation contravening the direction seen in the two

correlational studies. We had no reason to think that social distance would moderate the size of this effect. A subtler and riskier prediction was that the expectation effect would be smaller than required by expected-value considerations. We induced expectations so strong that a strictly value-maximizing person would either defect (if expectation of other's volunteering is high) or volunteer (if expectation is low). We doubted that these floors and ceilings would be empirically matched in size. Critically, we predicted that the shortfall relative to the floor of no volunteering would be greater than the shortfall relative to the ceiling of full volunteering. Such an asymmetry would constitute evidence of over-volunteering.

Method

We recruited 296 residents of the United States on Amazon Mechanical Turk and collected no further demographic information. Each participant received a small payment of c75 and a lottery ticket for a $25 Amazon.com gift card. Each participant responded to all four scenarios of the 2 (social distance: high vs. low) by 2 (expectation: high vs. low) design.[3]

The structure of the VoD and the social distance scale were introduced as in the previous studies, using a standard platform (Qualtrics Research Suite [Survey software], 2014). Participants were asked to consider only the closest (distance rank 1) and the remotest person as a partner in the VoD (social distance rank 100). For each dilemma, they were to assume either that this person was very likely to volunteer (with a 80% chance) or very unlikely to volunteer (20% chance). The order of the four scenarios was randomized over participants in a 2 (distant or closest partner first) × 2 (for the first partner: high or low expectations first) × 2 (for the second partner: high or low expectations first) design. Participants then entered their own likelihood to volunteer using a percentage scale.

Results and Discussion

Figure 4 shows the findings as bean plots with means and confidence intervals. Visual inspection reveals clear evidence for both the social distance hypothesis and the expectation hypothesis. We again used linear mixed models with random effects and bootstrapped confidence intervals for statistical analysis and effect-coding (−0.5 and 0.5) for the predictor variables social distance and expectation. The main effect of social distance, $b = -14.13$, 95% CI [−10.84, −17.54], $R^2 = 3\%$, indicated that participants were approximately 14% less likely to volunteer for the distant other compared with the close other. The main effect of expectation, $b = -23.30$, 95% CI [−18.52, −28.08], $R^2 = 9\%$, indicated that participants were about 23% less likely to volunteer when they expected the other to volunteer with a probability of 80% vs. 20%. The interaction term was not significant, $b = -2.92$, 95% CI [−8.64, 2.72], $R^2 < 0.1\%$.

The data also support the over-volunteering hypothesis. When the other was expected to volunteer with an 80% probability, the optimal response was to not volunteer at all. Yet, participants announced that they would volunteer with a 51 and 35% probability, respectively, for the close and distant other (see **Figure 4**). This is *prima facie* evidence for over-volunteering. Yet, there was also the converse effect of undervolunteering when the other was expected to volunteer with a 20% probability. Although the optimal response was to volunteer with certainty, participants announced that they would volunteer with a 72 and 60% probability, respectively, for the close and distant other.

The results of Study 3 replicate and extend the body of correlational findings accumulated in Studies 1 and 2. The social distance effect on volunteering is robust, consistent with the ideas of inclusive fitness (Hamilton, 1964) and strong reciprocity (Gintis, 2000). As the social distance heuristic uses a single cue, it opens the door to predictable error. We have identified over-volunteering as one such an error and we saw that respondents violate their own expectations regarding the choices of others when they arguably care the most about an efficient outcome. Study 3 shows that this violation of expectation occurs not only when these expectations are self-generated but also when they are externally provided.

GENERAL DISCUSSION

Summary and Review

Volunteer's Dilemmas pervade social life, although they are rarely recognized as such. Who will buy the wine for dinner? Who will start work on the co-authored manuscript? Who will punish the loafers and jaded bystanders (Przepiorka and Diekmann, 2013)? The VoD has received little research attention apart from the specific issue of bystander intervention and apathy in emergency situations (Darley and Latané, 1968; Krueger and Massey, 2009; Fischer et al., 2011). We suspect that the VoD is neglected because of the belief that it is easily resolved with a little goodwill and coordination, particularly among kin and the well-acquainted (Sir Karl Popper dissenting). Most research remains focused on social cooperation in public-goods and resource dilemmas involving unrelated strangers (Dawes, 1980; Norenzayan et al., 2016). In those dilemmas, collective outcomes continue to improve as more individuals contribute. In contrast, the relationship between collective welfare and the frequency of prosocial behavior is non-linear in the VoD. It is inefficient to have more than one volunteer or to have none at all. This non-linearity poses a psychological challenge. A prosocial person must consider the risk of making a redundant and thus inefficient contribution.[4]

An excess of prosociality can occur when individuals are close and when the effects of volunteering or mutual failure to volunteer are negative. Our principal explanation of this finding is the idea that people use a social-distance heuristic when deciding whether to accept the cost of volunteering. They are willing to make a sacrifice to the extent that the other person is socially, psychologically, or genetically

[3]An additional manipulation asked respondents to either seek to maximize their own payoffs or to maximize the joined payoffs. This manipulation had no effect on the results and is henceforth ignored.

[4]The VoD is akin to a step-level public-good dilemma, in which a benefit is provided to all once a threshold of contributive cooperation is reached. All additional contributions are wasted.

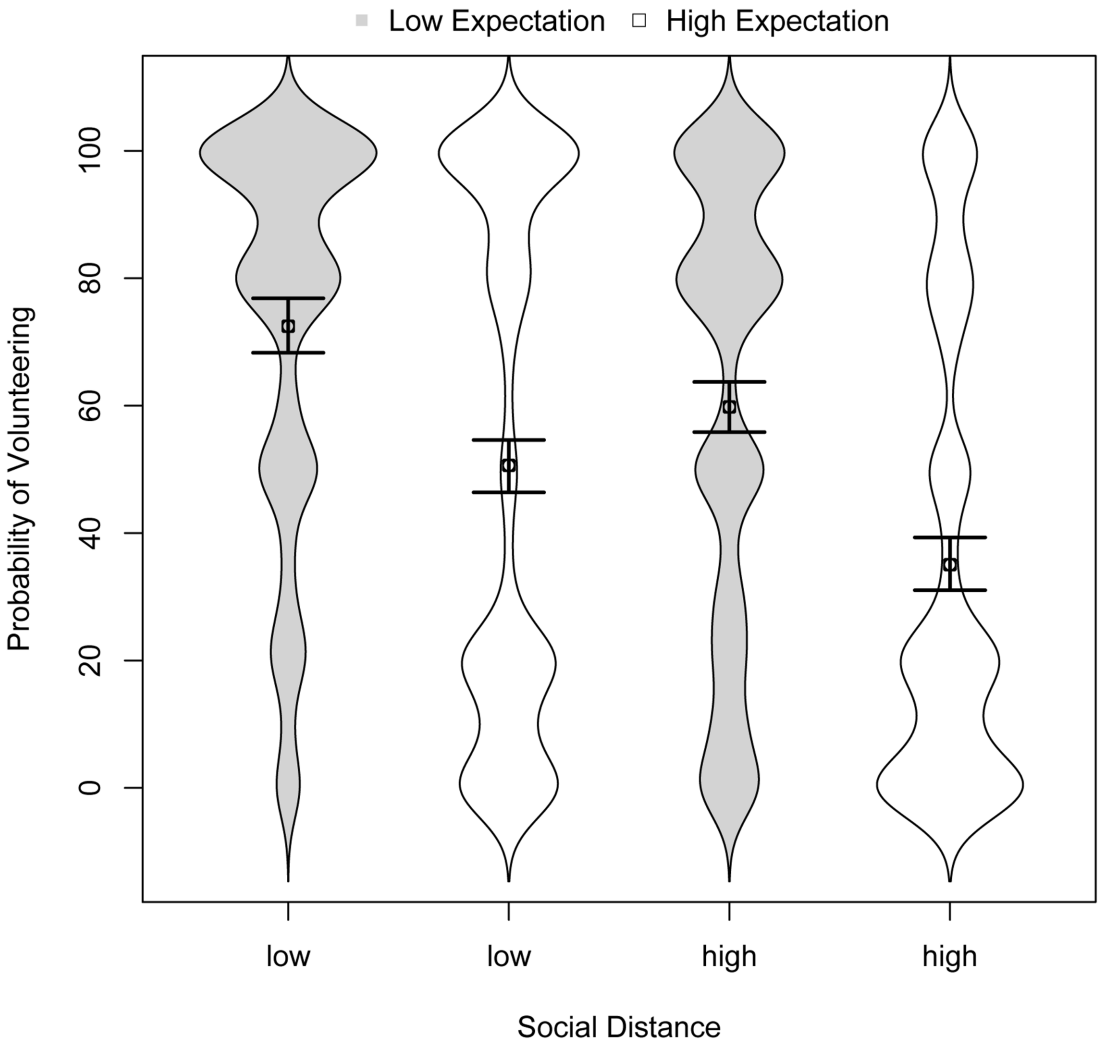


FIGURE 4 | Distributions of volunteering across conditions in Study 3 (and 95% confidence intervals for the means).

close to them. This heuristic works well in many contexts of interdependence, providing adaptive advantages that are recognized by evolutionary biologists and game theorists (e.g., Ferrière and Michod, 2011; Locey et al., 2013). Indeed, we find that the mean probability of volunteering tracks the predictions of a formal equilibrium model, which uses relatedness to weight and integrate the other person's outcomes with one's own (Archetti, 2009). When social distance is zero, the model assumes that players care for the outcomes of the other player as much as they do for their own.

Our findings suggest that many pairs of close individuals will end up with the same outcome, the R payoff for mutual volunteering, although they would have fared better if their probability of volunteering had been lower. It is not clear yet whether this effect is large enough so that individuals can gain insight into its non-optimality. Perhaps they will focus instead on the equality of their two payoffs, consider it fair, and find reassurance in the successful avoidance of the most aversive outcome of mutual defection (Leliveld et al., 2009). Alternatively, our findings point toward a mistaken sense of altruism (Krueger, 2011; Oakley et al., 2011), which, under certain conditions, can do great harm. For instance, when individual and group identities fuse, the eagerness to act prosocially can beget tragedy (Whitehouse et al., 2014).

Now consider the relevance of the findings regarding expectations of volunteering. With pain at stake, people expect close others to volunteer, and even over-volunteer. Why do respondents not scale back their own probability of volunteering to restore maximum efficiency? The logic of social projection suggests an answer (Krueger, 2013). Consider a person who is ready to volunteer and who expects others to do the same. This person cannot switch from 'volunteer' to 'defect' without assuming that others will do the same. If projection is a valid heuristic for inferring the actions of others, it is valid regardless of one's particular strategy. Like prosocial behavior, social projection decreases over social distance (Robbins and Krueger, 2005); this

general finding emerges in the present data too (Studies 1 and 2) and thus helps explain the tenacity of over-volunteering among close individuals.

If – as we believe – respondents generated their expectations about the likely behavior of others after they had made their own decisions, we can make sense of a final finding: respondents thought that the probability of others to volunteer was lower than their own. With volunteering being a socially desirable act, declaring oneself to be more willing to volunteer than others amounts to a better-than-average-effect (Alicke and Sedikides, 2009). Self-enhancers claim dual moral credit (Heck and Krueger, 2016). They not only volunteer but also predict that they volunteer more than others do. Self-enhancement is consistent with the general projective pattern (Heck and Krueger, 2015). If respondents derive expectations about others from their own decisions, these expectations should be more regressive (i.e., less extreme) than own decisions (Moore and Healy, 2008). Indeed, expectations were overall closer to the 50% mark than were judgments of own intended volunteering.

In light of the bounded rationality with which people approach the VoD, we may ask what options exist for efficient solutions. In contrast to the prisoner's dilemma and the assurance game, but like the game of chicken (Van Lange et al., 2014), the VoD yields best results if the two players act differently. Over repeated encounters, turn-taking in volunteering yields mutual benefits. In a one-shot episode, however, communication is of little help. If both individuals declare their intention to volunteer (or defect), additional factors must be brought in to break the tie. One reasonable social rule is to put the burden of volunteering on whomever can afford it the most (Przepiorka and Diekmann, 2013). When Linda and Laura reach for the lunch bill, jobless Linda may yield to working Laura (Abele et al., 2014). When there is no difference in wealth, timing is critical. Whoever announces their decision first forces the other to do the opposite (Schelling, 1960). We suspect that in such a sequential arrangement social distance will remain a moderating factor.

Open Questions

Our study designs reflect choices made under constraints and in the interest of expediency. Future research needs to identify and test pinpoint hypotheses to sharpen our theoretical understanding of the volunteer's dilemma and to enhance the generalizability of the findings.

First, there is the finding that over-volunteering occurred only for aversive outcomes. It may be too soon to declare valence a robust moderator as we had only one sample with a positive game frame. If, however, the valence effect survives further testing, we may note that the departure from rationality and adpativeness occurred where participants would arguably be most motivated to avoid it: in the domain of pain (Kahneman and Tversky, 1984; Baumeister et al., 2001).

Second, the task of mapping the effects of social distance onto the predictions of a rational equilibrium model limited us to a artificial methodology. To scale social distance with precision, we sacrificed the real-life experience of encountering others in the dilemma. As future research meets the challenge of mundane realism, it will be critical to remain wary of confounds. Individuated partners will introduce a host of additional information or assumptions that might increase the variability of results in random or systematic ways.

Third, the use of five levels of social distance presented in non-random order may raise the specter of experimental demand. Yet, we remain sanguine because the demand hypothesis makes no specific predictions. What particular slope or which specific intercept, for example, should a respondent feel called upon to produce when scaling her own willingness to volunteer onto social distance?

Fourth, we presented the VoD as a choice problem of the type used in scenario research in the psychology of judgment and decision-making (see, for example, Murnighan et al., 1993, or Kim and Murnighan, 1997, for such work on the VoD). In contrast, behavioral economics prizes consumable payoffs. Recent work in our laboratory suggests that in the VoD, symbolic payoffs yield the same results as material ones do (Krueger et al., 2016, unpublished).

Many ordinary people and the scientists who study them operate from the simple, reasonable, and adaptive heuristic that prosocial behavior is socially desirable. Their moral concerns take the form of asking what can be done to make such behavior more common. Our excursion into the volunteer's dilemma suggests structural and psychological factors can combine to undercut the effects of good intentions and expectations. More is not always better.

ETHICS STATEMENT

The studies were exempt. Survey research with no conceivable risk to participants.

AUTHOR CONTRIBUTIONS

JK conceived the project, supervised data collection, consulted with data analysis, and drafted the manuscript. JU conceived the project, supervised data collection, consulted with data analysis, and drafted the manuscript. LC conceived the third experiment, collected and analyzed data, and helped with manuscript preparation.

ACKNOWLEDGMENTS

We thank the volunteers. Marco Archetti provided helpful comments on a previous version of this manuscript. Carolin Strobl and Michel Philipp provided helpful comments on the statistical analyses. Tony Evans, Gideon Goldin, Angela Gross, Anna Hartley, Patrick Heck, and Nicolas Ramer helped collect and manage data. Data, code, and materials are archived at https://osf.io/26eag/.

REFERENCES

Abele, S., Stasser, G., and Chartier, C. (2014). Use of social knowledge in tacit coordination: social focal points. *Organ. Behav. Hum. Decis. Process.* 123, 23–33. doi: 10.1016/j.obhdp.2013.10.005

Ajzen, I., and Fishbein, M. (2008). Scaling and testing multiplicative combinations in the expectancy-value model of attitudes. *J. Appl. Soc. Psychol.* 38, 2222–2247. doi: 10.1111/j.1559-1816.2008.00389.x

Alicke, M. D., and Sedikides, C. (2009). Self-enhancement and self-protection: what they are and what they do. *Eur. Rev. Soc. Psychol.* 20, 1–48. doi: 10.1080/10463280802613866

Alpern, S., and Reyniers, D. (2001). Games of crowding. *Int. Game Theory Rev.* 3, 27–56. doi: 10.1142/S0219198901000294

Archetti, M. (2009). The volunteer's dilemma and the optimal size of a social group. *J. Theor. Biol.* 261, 475–480. doi: 10.1016/j.jtbi.2009.08.018

Archetti, M., and Scheuring, I. (2011). Coexistence of cooperation and defection in public goods games. *Evolution* 65, 1140–1148. doi: 10.1111/j.1558-5646.2010.01185.x

Balliet, D., Wu, J., and De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychol. Bull.* 140, 1556–1581. doi: 10.1037/a0037737

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-7*. Available at: http://CRAN.Rproject.org/package = lme4

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Rev. Gen. Psychol.* 5, 323–370. doi: 10.1037/1089-2680.5.4.323

Berg, N., and Gigerenzer, G. (2010). As-if behavioral economics: neo-classical economics in disguise? *Hist. Econ. Ideas* 18, 133–165.

Bernoulli, D. (1954/1738). Exposition of a new theory on the measurement of risk". *Econometrica* 22, 23–36. doi: 10.2307/1909829

Binmore, K. (2007). *Game Theory: A Very Short Introduction*. Oxford: Oxford University Press.

Brewer, M. B. (2008). "Depersonalized trust and ingroup cooperation," in *Rationality and Social Responsibility: Essays in Honor of Robyn Mason Dawes*, ed. J. I. Krueger (New York, NY: Psychology Press), 215–232.

Burnstein, E., Crandall, C., and Kitayama, S. (1994). Some neo-Darwinian decision rules for altruism: weighing cues for inclusive fitness as a function of the biological importance of the decision. *J. Personal. Soc. Psychol.* 67, 773–789. doi: 10.1037/0022-3514.67.5.773

Buss, D. M. (1999). *Evolutionary Psychology*. Boston, MA: Allyn and Bacon.

Darley, J. M., and Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *J. Personal. Soc. Psychol.* 8, 377–383. doi: 10.1037/h0025589

Dawes, R. M. (1980). Social dilemmas. *Annu. Rev. Psychol.* 31, 169–193. doi: 10.1146/annurev.ps.31.020180.001125

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *J. Exp. Soc. Psychol.* 25, 1–17. doi: 10.1016/0022-1031(89)90036-X

Dawes, R. M., McTavish, T., and Shacklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *J. Personal. Soc. Psychol.* 35, 1–11. doi: 10.1037/0022-3514.35.1.1

Dawkins, R. (1976). *The Selfish Gene*. New York, NY: Oxford University Press.

De Dreu, C. K. W., Dussel, D. B., and Ten Velden, F. S. (2015). In intergroup conflict, self-sacrifice is stronger among pro-social individuals, and parochial altruism emerges especially among cognitively taxed individuals. *Front. Psychol.* 6:572. doi: 10.3389/fpsyg.2015.00572

DiDonato, T. E., Ullrich, J., and Krueger, J. I. (2011). Social perception as induction and inference: an integrative model of intergroup differentiation, ingroup favoritism, and differential accuracy. *J. Pers. Soc. Psychol.* 100, 66–83. doi: 10.1037/a0021051

Diekmann, A. (1985). Volunteer's dilemma. *J. Conflict Resolut.* 29, 605–610. doi: 10.1177/0022002785029004003

Fehr, E., Fischbacher, U., and Kosfeld, M. (2005). Neuroeconomic foundations of trust and social preferences: Initial evidence. *Am. Econ. Rev.* 95, 346–351. doi: 10.1257/000282805774669736

Ferrière, R., and Michod, R. E. (2011). Inclusive fitness in evolution. *Nature* 471, E6–E7. doi: 10.1038/nature09834

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* 71, 397–404. doi: 10.1016/S0165-1765(01)00394-9

Fischer, I. (2009). Friend or foe: Subjective expected relative similarity as a determinant of cooperation. *J. Exp. Psychol. Gen.* 138, 341–350. doi: 10.1037/a0016073

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrinic, C., Kastenmüler, A., Frey, D., et al. (2011). The bystander effect: a meta-analytic review on bystander intervention in dangerous and non- dangerous emergencies. *Psychol. Bull.* 137, 517–537. doi: 10.1037/a0023304

Gaschler, R., Schwager, S., Umbach, V. J., Frensch, P. A., and Schubert, T. (2014). Expectation mismatch: differences between self-generated and cue-induced expectations. *Neurosci. Biobehav. Rev.* 46, 139–157. doi: 10.1016/j.neubiorev.2014.06.009

Gigerenzer, G., and Gaissmaier, W. (2011). Heuristic decision making. *Annu. Rev. Psychol.* 62, 451–482. doi: 10.1146/annurev-psych-120709-145346

Gintis, H. (2000). Strong reciprocity and human sociality. *J. Theor. Biol.* 206, 169–179. doi: 10.1006/jtbi.2000.2111

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *J. Theor. Biol.* 7, 1–16. doi: 10.1016/0022-5193(64)90038-4

Hardin, G. (1968). The tragedy of the commons. *Science* 162, 243–248.

Heck, P. R., and Krueger, J. I. (2015). Self-enhancement diminished. *J. Exp. Psychol. Gen.* 144, 1003–1020. doi: 10.1037/xge0000105

Heck, P. R., and Krueger, J. I. (2016). Social perception of self-enhancement bias and error. *Soc. Psychol.* 2016, 1–13. doi: 10.1027/1864-9335/a000287

Hoffrage, U., and Hertwig, R. (2012). "Simple heuristics in a complex social world," in *Social Judgment and Decision Making*, ed. J. I. Krueger (New York, NY: Psychology Press), 135–150.

Jones, B. A., and Rachlin, H. (2006). Social discounting. *Psychol. Sci.* 17, 283–286. doi: 10.1111/j.1467-9280.2006.01699.x

Kahneman, D., and Tversky, A. (1984). Choices, values, and frames. *Am. Psychol.* 39, 341–350. doi: 10.1037/0003-066X.39.4.341

Kampstra, P. (2008). Beanplot: a boxplot alternative for visual comparison of distributions. *J. Stat. Softw.* 28, 1–9.

Kim, J. W., and Murnighan, J. K. (1997). The effects of connectedness and self-interest in the organizational volunteer dilemma. *Int. J. Confl. Manag.* 8, 32–51. doi: 10.1108/eb022789

Krueger, J., and Clement, R. W. (1994). The truly false consensus effect: an ineradicable and egocentric bias in social perception. *J. Pers. Soc. Psychol.* 67, 596–610. doi: 10.1037/0022-3514.67.4.596

Krueger, J. I. (2003). Return of the ego—self-referent information as a filter for social prediction: comment on Karniol (2003). *Psychol. Rev.* 110, 585–590. doi: 10.1037/0033-295X.110.3.585

Krueger, J. I. (2007). From social projection to social behaviour. *Eur. Rev. Soc. Psychol.* 18, 1–35. doi: 10.1080/10463280701284645

Krueger, J. I. (2011). "Altruism gone mad," in *Pathological Altruism*, eds B. Oakley, A. Knafo, G. Madhavan, and D. S. Wilson (New York, NY: Oxford University Press), 392–402.

Krueger, J. I. (2013). Social projection as a source of cooperation. *Curr. Dir. Psychol. Sci.* 22, 289–294. doi: 10.1177/0963721413481352

Krueger, J. I. (2014). Heuristic game theory. *Decision* 1, 59–61. doi: 10.1037/dec0000002

Krueger, J. I., and Acevedo, M. (2005). "Social projection and the psychology of choice," in *The Self in Social Perception*, eds M. D. Alicke, D. Dunning, and J. I. Krueger (New York, NY: Psychology Press), 17–41.

Krueger, J. I., and DiDonato, T. E. (2008). Social categorization and the perception of groups and group differences. *Soc. Personal. Psychol. Compass* 2, 733–750. doi: 10.1111/j.1751-9004.2008.00083.x

Krueger, J. I., DiDonato, T. E., and Freestone, D. (2012). Social projection can solve social dilemmas. *Psychol. Inq.* 23, 1–27. doi: 10.1080/1047840X.2012.641167

Krueger, J. I., Evans, A. M., and Heck, P. R. (2016). "Let me help you help me: trust between profit and prosociality," in *Social Dilemmas: New Perspectives on Trust*, eds P. A. M. Van Lange, B. Rockenbach, and T. Yamagishi (New York, NY: Oxford University Press).

Krueger, J. I., and Massey, A. L. (2009). A rational reconstruction of misbehavior. *Soc. Cogn.* 27, 785–810. doi: 10.1521/soco.2009.27.5.786

Kurzban, R., Burton-Chellew, M. N., and West, S. A. (2015). The evolution of altruism in humans. *Annu. Rev. Psychol.* 66, 575–599. doi: 10.1146/annurev-psych-010814-015355

LaHuis, D. M., Hartman, M. J., Hakoyama, S., and Clark, P. C. (2014). Explained variance measures for multilevel models. *Organ. Res. Methods* 17, 433–451. doi: 10.1177/1094428114541701

Leliveld, M., Van Beest, I., Van Dijk, E., and Tenbrunsel, A. E. (2009). Understanding the influence of outcome valence in bargaining: a study on fairness accessibility, norms, and behavior. *J. Exp. Soc. Psychol.* 45, 505–514. doi: 10.1016/j.jesp.2009.02.006

Locey, M. L., Safin, V., and Rachlin, H. (2013). Social discounting and the prisoner's dilemma game. *J. Exp. Anal. Behav.* 99, 85–97. doi: 10.1002/jeab.3

Luce, R. D., and Raiffa, H. (1957). *Games and Decisions*. New York, NY: Wiley.

Messé, L., and Sivacek, J. M. (1979). Predictions of others' responses in a mixed-motive game: self-justification or false consensus? *J. Personal. Soc. Psychol.* 37, 602–607. doi: 10.1037/0022-3514.37.4.602

Monterosso, J., and Ainslee, G. (2003). Game theory need not abandon individual maximization. *Behav. and Brain Sci.* 26:171. doi: 10.1017/S0140525X03400058

Moore, D. A., and Healy, P. J. (2008). The trouble with overconfidence. *Psychol. Rev.* 115, 502–517. doi: 10.1037/0033-295X.115.2.502

Murnighan, J. K., Kim, J. W., and Metzger, A. R. (1993). The volunteer dilemma. *Adm. Sci. Q.* 38, 515–538. doi: 10.2307/2393335

Murphy, R. O., and Ackermann, K. A. (2014). Social value orientation: theoretical and measurement issues in the study of social preferences. *Pers. Soc. Psychol. Rev.* 18, 13–41. doi: 10.1177/1088868313501745

Nielsen, U. H., Tyran, J.-R., and Wengström, E. (2014). Second thoughts on free riding. *Econ. Lett.* 122, 136–139. doi: 10.1007/s10728-012-0221-4

Norenzayan, A., Shariff, A. F., Willard, A. K., Slingerland, E., Gervais, W. M., McNamara, R. A., et al. (2016). The cultural evolution of prosocial religions. *Behav. Brain Sci.* 39, 1–19. doi: 10.1017/S0140525X14001356

Oakley, B., Knafo, A., Madhavan, G., and Wilson, D. S. (2011). *Pathological Altruism*. New York, NY: Oxford University Press.

Pascal, B. (1995/1669). *Pensées*, ed. A. J. Krailsheimer (London: Penguin Books).

Popper, K. (1945/2011). *The Open Society and its Enemies*. London: Routledge.

Pruitt, D. G., and Kimmel, M. J. (1977). Twenty years of experimental gaming: critique, synthesis, and suggestions for the future. *Annu. Rev. Psychol.* 28, 363–392. doi: 10.1146/annurev.ps.28.020177.002051

Przepiorka, W., and Diekmann, A. (2013). Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proc. R. Soc. B* 280, 20130247. doi: 10.1098/rspb.2013.0247

Qualtrics Research Suite [Survey software] (2014). *Qualtrics Research Suite [Survey software]*. Provo, UT: Qualtrics.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabbie, J. M., and Horwitz, M. (1969). Arousal of ingroup-outgroup bias by a chance win or loss. *J. Pers. Soc. Psychol.* 13, 269–277. doi: 10.1037/h0028284

Rapoport, A. (1967). A note on the index of cooperation for Prisoner's Dilemma. *J. Confl. Resolut.* 11, 101–103. doi: 10.1177/002200276701100108

Rapoport, A. (2003). Chance, utility, rationality, equilibrium. *Behav. Brain Sci.* 26, 172–173. doi: 10.1017/S0140525X03420050

Rapoport, A., and Chammah, A. M. (1966). The game of chicken. *Am. Behav. Sci.* 10, 10–28. doi: 10.1177/000276426601000303

Robbins, J. M., and Krueger, J. I. (2005). Social projection to ingroups and outgroups: a review and meta-analysis. *Pers. Soc. Psychol. Rev.* 9, 32–47. doi: 10.1207/s15327957pspr0901_3

Schelling, T. C. (1960). *Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17, 551–566. doi: 10.1037/a0029487

Tajfel, H., and Turner, J. (1979). "An integrative theory of intergroup conflict," in *The Social Psychology of Intergroup Relations*, eds W. G. Austin and S. Worchel (Monterey, CA: Brooks/Cole), 33–48.

Thomas, K. A., DeScioli, P., Haque, O. S., and Pinker, S. (2014). The psychology of coordination and common knowledge. *J. Personal. Soc. Psychol.* 107, 657–676. doi: 10.1037/a0037037

Tversky, A., and Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychol. Sci.* 3, 305–309. doi: 10.1111/j.1467-9280.1992.tb00678.x

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: an integrative model of social value orientation. *J. Personal. Soc. Psychol.* 77, 337–349. doi: 10.1016/j.actpsy.2008.01.005

Van Lange, P. A. M., Balliet, D., Parks, C. D., and Van Vugt, M. (2014). *Social Dilemmas: The Psychology of Human Cooperation*. New York, NY: Oxford University Press.

Van Veelen, R., Otten, S., Cadinu, M. R., and Hansen, N. (2016). An integrative model of social identification: self-stereotyping and self-anchoring as two cognitive pathways. *Pers. Soc. Psychol. Rev.* 23, 2–36. doi: 10.1177/1088868315576642

Von Neumann, J., and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Whitehouse, H., McQuinn, B., Buhrmester, M., and Swann, W. B. Jr. (2014). Brothers in arms: Lybian revolutionaries bond like family. *Proc. Natl. Acad. Sci. U.S.A.* 111, 17783–17785. doi: 10.1073/pnas.1416284111

Yamagishi, T., and Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Soc. Psychol. Q.* 63, 116–132. doi: 10.2307/2695887

Conflict of Interest Statement: 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.

Copyright © 2016 Krueger, Ullrich and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 August 2016
doi: 10.3389/fpsyg.2016.01312

Cognitive Load Does Not Affect the Behavioral and Cognitive Foundations of Social Cooperation

Laura Mieth, Raoul Bell and Axel Buchner*

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

The present study serves to test whether the cognitive mechanisms underlying social cooperation are affected by cognitive load. Participants interacted with trustworthy-looking and untrustworthy-looking partners in a sequential Prisoner's Dilemma Game. Facial trustworthiness was manipulated to stimulate expectations about the future behavior of the partners which were either violated or confirmed by the partners' cheating or cooperation during the game. In a source memory test, participants were required to recognize the partners and to classify them as cheaters or cooperators. A multinomial model was used to disentangle item memory, source memory and guessing processes. We found an expectancy-congruent bias toward guessing that trustworthy-looking partners were more likely to be associated with cooperation than untrustworthy-looking partners. Source memory was enhanced for cheating that violated the participants' positive expectations about trustworthy-looking partners. We were interested in whether or not this expectancy-violation effect—that helps to revise unjustified expectations about trustworthy-looking partners—depends on cognitive load induced via a secondary continuous reaction time task. Although this secondary task interfered with working memory processes in a validation study, both the expectancy-congruent guessing bias as well as the expectancy-violation effect were obtained with and without cognitive load. These findings support the hypothesis that the expectancy-violation effect is due to a simple mechanism that does not rely on demanding elaborative processes. We conclude that most cognitive mechanisms underlying social cooperation presumably operate automatically so that they remain unaffected by cognitive load.

Keywords: dual task, working memory load, trust, social cooperation, source memory

OPEN ACCESS

Edited by:
Anna Thorwart,
University of Marburg, Germany

Reviewed by:
Diane Swick,
VA Northern California Health Care System, USA
Brittany S. Cassidy,
Indiana University Bloomington, USA
Danielle M. Shore,
University of Oxford, UK

****Correspondence:***
Laura Mieth
Laura.Mieth@hhu.de
Raoul Bell
Raoul.Bell@hhu.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 May 2016*
Accepted: *17 August 2016*
Published: *31 August 2016*

Citation:
Mieth L, Bell R and Buchner A (2016) Cognitive Load Does Not Affect the Behavioral and Cognitive Foundations of Social Cooperation. Front. Psychol. 7:1312. doi: 10.3389/fpsyg.2016.01312

INTRODUCTION

There is increasing interest in whether (and how) social cooperation is affected by cognitive load. Although it has been proposed that cooperation is generally decreased (Piovesan and Wengström, 2009) or enhanced (Rand et al., 2012) by cognitive load, no consensus about this issue has been reached, and there are a number of null findings and failed replications (Tinghög et al., 2013; Kessler and Meier, 2014; Verkoeijen and Bouwmeester, 2014). Focusing on how cognitive load

affects specific cognitive mechanisms that are important for cooperation could be a more promising approach than looking at the global outcome of presumably many different kinds of processes involved in cooperation. Therefore, the present study examines how memory for cheating or cooperation—a necessary prerequisite for reciprocal cooperation (Trivers, 1971)—is affected by cognitive load. We were particularly interested in whether or not social expectations affect the participants' memory for the cheating or cooperation of interaction partners under cognitive load.

Examining the influence of social expectations seems particularly important because social cooperation depends fundamentally on expectations about other people's behaviors. This can be illustrated with the Prisoner's Dilemma Game (Clark and Sefton, 2001), which serves as a model for understanding human cooperation. In this game, two players independently decide whether or not to cooperate with each other. Mutual cooperation leads to reward while mutual defection leads to punishment, which reflects that more can be achieved through cooperation. However, unilateral defection leads to the highest payoff (the temptation payoff) while unilateral cooperation leads to the worst payoff (the sucker's payoff). The dilemma lies in the fact that each player can maximize his or her payoff by defecting, but mutual defection leads to a worse payoff for both players than mutual cooperation. Humans are often able to resist the selfish temptation to defect, and high levels of cooperation are often achieved even in one-shot games (Delton et al., 2011). However, given that nobody wants to be suckered, cooperation depends on people's expectations about whether or not the other player will choose to cooperate.

These expectations are strongly influenced by facial appearance (Chang et al., 2010; Olivola and Todorov, 2010). Appearance-based impressions are formed quickly (Willis and Todorov, 2006; Todorov et al., 2009) and automatically (Engell et al., 2007), but are quite stable over time. There is also a high degree of inter-individual agreement about who looks trustworthy and who does not (Todorov, 2008). These appearance-based impressions determine people's behaviors in social-dilemma games: People often cooperate with trustworthy-looking partners, and defect against untrustworthy-looking partners (van 't Wout and Sanfey, 2008; Rezlescu et al., 2012).

However, appearance-based expectations may often turn out to be false. People are somewhat better than chance when using facial appearance to predict whether partners will cooperate or cheat in social-dilemma games (Bonnefon et al., 2013), but facial appearance is a comparatively invalid source of information about a person's character, and people rely on it more than they should (Olivola and Todorov, 2010). Therefore, remembering expectancy-incongruent information is especially important to correct invalid appearance-based impressions about other persons. To correct a false impression, it is insufficient to simply recognize the face as familiar, it is also necessary to have good source memory for the association between the face and the behavior of the person (Buchner et al., 2009). For example, remembering that a trustworthy-looking person is unreliable is important to avoid being misled by the person's trustworthy appearance in the future. This functional analysis leads to the prediction that people should have better source memory for expectancy-incongruent information than for expectancy-congruent information.

The same prediction can be derived from schema theories of memory. The schema-copy-plus-tag model (Graesser and Nakamura, 1982) implies that expectancy-congruent behaviors are represented in memory by pointers to general schemas. Expectancy-violating behaviors are tagged as schema violations. In memory tests, participants often produce a high amount of schema-congruent information due to guessing, but memory accuracy is often poor for this type of information because it is produced regardless of whether it was present at encoding or not. The discrimination between actually experienced and new information is often better for schema-atypical information. For instance, participants will guess that a trustworthy-looking face belongs to a trustworthy person, regardless of whether the behavior of the person was trustworthy or not. Learning that a trustworthy-looking person is a cheater represents a more distinct and therefore more memorable information. Indeed, several studies confirmed the idea that people remember appearance-incongruent behaviors better than appearance-congruent behaviors (Suzuki and Suga, 2010; Volstorf et al., 2011; Bell et al., 2012b).

The present study serves to test whether or not the memory advantage for expectancy-incongruent behavior depends on cognitive load. Two opposing hypotheses are tested. Source memory for cheating and cooperation may be *impaired* by cognitive load because source memory is often believed to be more fragile and more dependent on cognitive resources than familiarity-based item memory (Nieznański, 2013). Therefore, the encoding of the association between a face and cheating or cooperation may be decreased under cognitive load. Memory for expectancy-incongruent information in particular may be negatively affected because this information cannot be easily integrated into existing schemas. Expectancy-incongruent information may trigger more effortful elaborative encoding than expectancy-congruent information, which will lead to enhanced memory for this information under normal circumstances. However, these elaborative processes may depend on the mobilization of additional cognitive resources. Therefore, a reduction in available cognitive resources may eliminate the expectancy-violation effect. Consistent with this hypothesis, the source memory advantage for expectancy-incongruent information was absent in older adults (Bell et al., 2013) who may have fewer cognitive resources available than younger adults. If the memory advantage for expectancy-incongruent information is abolished under cognitive load, our ability to successfully engage in social cooperation would be impaired because this type of memory is essential for correcting maladaptive behavior tendencies.

However, it is also possible that cognitive load has *no effect* on memory for expectancy-incongruent behaviors. Remembering expectancy-incongruent information seems to be too important to vanish quickly under conditions of high cognitive load. Cooperation is particularly important in stressful situations. The human cognitive system would be badly designed if it would

let go of the most important information under distracting and stressful conditions first. Therefore, the cognitive machinery specialized in categorizing other people are often assumed to be automatic (Klein et al., 2002). The same hypothesis can be based on non-functional, schema-based accounts of memory. According to the schema-copy-plus-tag model, schema-atypical information is encoded and retained in the form of unelaborated tags. This encoding strategy is assumed to be frugal in terms of processing resources, and should remain unaffected by cognitive load (Graesser and Nakamura, 1982). Accordingly, source memory for the face of a cheater is often not due to an enhanced recollection of the specific details of the cheating episode, but instead due to the rough classification of the person as a "cheater" in form of emotional tagging (Bell et al., 2012a). Arguably, these unelaborated emotional tags can be automatically encoded even under conditions of high cognitive load. Consistent with this idea, a demanding secondary task at encoding does not always lead to decreased memory for schema-atypical information, but may even result in a more pronounced schema-atypicality effect in source memory (Ehrenberg and Klauer, 2005). The automatic tagging of expectancy-violating behaviors would allow people to successfully engage in social cooperation even under stressful and distracting conditions.

The present series of experiments was designed to discriminate between these two conflicting hypotheses. The first experiment served to replicate the finding that source memory for the cheating or cooperation of others is enhanced for appearance-incongruent behaviors. To anticipate, an asymmetrical source memory advantage for appearance-incongruent cheating was found. In two further experiments, we examined whether this incongruity advantage would vanish under conditions of increased cognitive load. A fourth study was designed to validate the cognitive-load task by showing that this task does indeed interfere with (general) working-memory resources.

EXPERIMENT 1

Experiment 1 served as a replication of the effects reported by Bell et al. (2012b) with the only difference that female instead of male faces were used as stimuli. We expected to replicate the finding that people guess that trustworthy-looking faces would be associated with cooperation and untrustworthy-looking faces with cheating. Furthermore, we expected that participants would remember appearance-incongruent behaviors better than appearance-congruent behaviors. In most experiments (Suzuki and Suga, 2010; Bell et al., 2012b), this memory advantage was asymmetric in that participants remembered cheating better than cooperation when the partners looked trustworthy, but there was only a non-significant tendency toward remembering cooperation better than cheating when the partners looked untrustworthy. This asymmetry should be particularly pronounced for female faces because they elicit more positive social expectations than male faces, which means that the violation of these positive expectations is particularly salient when female faces are used (Kroneisen and Bell, 2013).

Method

Participants

One hundred and twelve students (73 of whom were female) with a mean age of 23 ($SD = 5$) participated in Experiment 1 (**Table 1**). All participants gave written informed consent in accordance with the Declaration of Helsinki. The present experiments are part of a series of experiments that has been approved by the ethics committee of the Department of Experimental Psychology at Heinrich Heine University Düsseldorf.

Materials, Procedure, and Design

The same sequential Prisoner's Dilemma Game was used as in previous studies (Bell et al., 2012b, 2013). In this game, participants were required to invest money into a joint business with partners whose faces were shown on the screen. Participants played with 20 trustworthy-looking partners and with 20 untrustworthy-looking partners. The faces were randomly drawn from a set of 40 trustworthy-looking and 40 untrustworthy-looking frontal facial photographs of women[1] with a neutral expression (250 × 375 pixel) from the FERET database (Phillips et al., 1998). In a norming study, the untrustworthy-looking faces had received low trustworthiness ratings ($M = 2.75$, $SD = 0.24$) and the trustworthy-looking faces had received high trustworthiness ratings ($M = 4.28$, $SD = 0.23$) on a scale ranging from 1 to 6. Half of the partners in each condition cooperated and the other half cheated.

Participants could familiarize themselves with the game in two practice trials. At the start of the game, they were informed that they played for real money. In each trial, participants first saw a silhouette at the left side of the screen (representing the participant), and the partner's face at the right side of the screen (**Figure 1**). Participants were required to decide whether to invest 15 cents or 30 cents (by pressing a left or right button of the response box, respectively). The decision was displayed on screen for 1 s. The investment was presented in an arrow for 500 ms before it moved to the center of the screen within 500 ms. Similarly, the partner's decision was shown in an arrow

[1]As in our previous studies (e.g., Bell et al., 2015) we only used faces from one gender because it is well known that female faces are more trustworthy than male faces (Kroneisen and Bell, 2013), and we did not want facial gender to dilute the facial trustworthiness manipulation.

TABLE 1 | Comparison of age, gender, and justice sensitivity (Schmitt et al., 2005) of Experiment 1 and 2 and Experiment 1 and 3, respectively.

	Age	Gender	Justice Sensitivity
Experiment 1	$M = 23$; $SD = 5$	female = 73 male = 39	$M = 2.93$; $SD = 0.60$
Experiment 2	$M = 24$; $SD = 5$	female = 67 male = 42	$M = 2.80$; $SD = 0.70$
Experiment 3	$M = 22$; $SD = 5$	female = 69 male = 34	—
Comparison of Experiment 1 and 2	$t(219) = 1.78$, $p = 0.08$	χ^2 (1) = 0.33, $p = 0.57$	$t(219) = 1.42$, $p = 0.16$
Comparison of Experiment 1 and 3	$t(213) = 1.09$, $p = 0.28$	χ^2 (1) = 0.08, $p = 0.78$	—

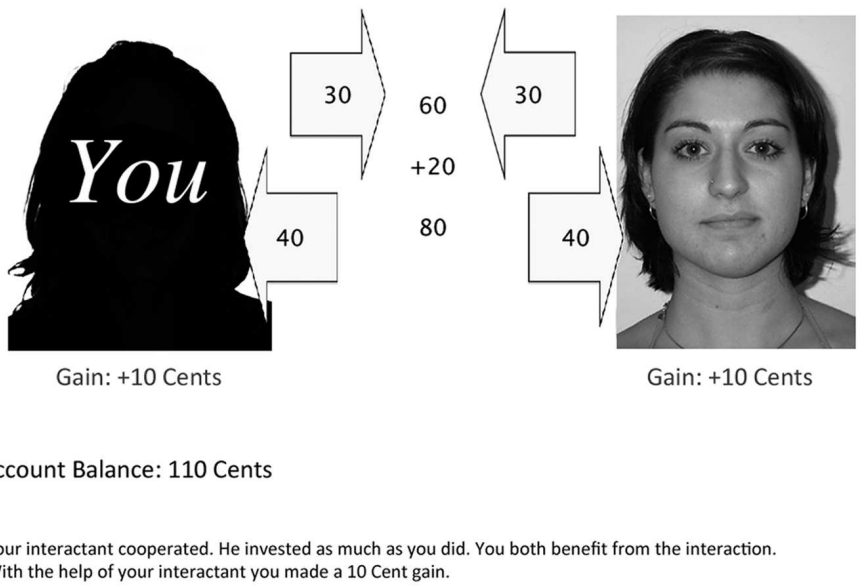


FIGURE 1 | A screenshot of the sequential prisoner's dilemma game. In this example, both the participant and the partner cooperated and invested 30 cents, resulting in a 10 cents gain for each of them. The partner's photograph shown in this example was taken from the Center for Vital Longevity (CVL) face database (Minear and Park, 2004).

for 500 ms, before it moved to the center of the screen within 500 ms. The sum of investments was then shown in the middle of the screen. After 500 ms a bonus of 1/3 of the sum of investments was added. After 500 ms, the total sum was shown. After a further 500 ms, this total sum was split up between the partners. Both the participant and the partner received half of the total sum, regardless of what they had invested. The partner's share was shown in an arrow moving toward the partner's face (500 ms). After 500 ms, the participant's share was shown in an arrow moving to the participant's silhouette (500 ms). After 1 s, the partner's gain or loss was presented, followed by the participant's gain or loss (after 500 ms). After a further 500 ms, the updated account balance of the participant was presented, and (again after 500 ms) a summary of the interaction was displayed. The next trial was initiated by the participant pressing the continue button.

A cooperating partner always reciprocated the participant's investment (either 15 or 30 cents), which resulted in a gain for both players. A cheating partner invested nothing (0 cents), which resulted in a gain for the partner at the expense of the participant, who lost money.

The payoff (gain or loss) of each player can be determined by the formula:

$$P_a = \frac{I_a + I_b + \frac{1}{3} \cdot (I_a + I_b)}{2} - I_a$$

where P_a is the payoff of Player A, I_a is the investment of Player A, and I_b is the investment of Player B. Applying this formula, it is obvious that interacting with a cooperating partner led to a gain, and interacting with a cheating partner led to a loss of the same magnitude for the participant.

After the game, participants received the instructions for the surprise source memory test. Eighty faces were presented. Half of the faces were old (presented during the sequential Prisoner's Dilemma Game), and the other half were new. Participants were first required to rate the likability of the faces on a scale ranging from 1 (not likable at all) to 6 (very likable). After pressing the continue button, participants were asked whether or not they had seen the face during the game. If participants indicated that they had seen the face before, they were required to decide whether the face belonged to a cheater or to a cooperator. After pressing the continue button, the next face was shown. Before leaving, participants filled out a paper–pencil version of the justice sensitivity questionnaire (Schmitt et al., 2005), and were paid.

The design was a 2 × 2 repeated measures design with facial trustworthiness (trustworthy vs. untrustworthy) and behavior (cheating vs. cooperation) as independent variables. Dependent variables were game investments, likability ratings, and memory performance. A multinomial model was used to distinguish among old–new recognition, source memory, and guessing processes. Given $\alpha = 0.05$, a sample size of $N = 112$, and 80 responses in the source memory test, it was possible to detect an effect of size $w = 0.04$ (comparable to the effect sizes observed by Buchner et al., 2009; Küppers and Bayen, 2014; Bell et al., 2015; Kroneisen et al., 2015) for the comparison between source

memory for cheaters and cooperators with a statistical power (1 – β) of 0.97. The power calculation was performed using G*Power (Faul et al., 2007).

Measuring Source Memory

When examining source memory, it is important to use a measure that does not confound item recognition, source memory, and guessing (Bröder and Meiser, 2007). Therefore, we applied the widely used (Erdfelder et al., 2009) source monitoring model of Bayen et al. (1996) to measure source memory and source guessing separately.

To illustrate, the first model tree in **Figure 2** represents the cognitive states that are assumed to underlie the classification of a cheater face. With probability D_{Cheat}, participants know that the face is old (remember that they have seen the face during the game). With probability d_{Cheat}, they also have source memory for the face (remember that the person is a cheater). The source memory parameter is expressed as a conditional probability that varies between 0 and 1. A probability of 0 represents the absence of source memory while a probability of 1 represents perfect source memory. If participants fail to remember the source, which occurs with the complementary probability 1 – d_{Cheat}, they may guess, with probability *g*, that the person was a cheater or, with probability 1 – *g*, that the person was a cooperator. If they fail to recognize the face as old, which occurs with probability 1 – D_{Cheat}, they may guess, with probability *b*, that the face is old, and may then guess that the person was a cheater with probability *g,* or that the person was a cooperator with probability 1 – *g*. With probability 1 – *b*, participants may guess that the face is new (has not been encountered during the game). The goodness-of-fit tests are based on the log-likelihood ratio statistic G^2 which is asymptotically chi-square distributed (Riefer and Batchelder, 1988; Stahl and Klauer, 2007; Singmann and Kellen, 2013). Parameter estimations and goodness-of-fit tests were calculated using multiTree (Moshagen, 2010). The observed response frequencies for Experiments 1–3 are reported in the Online Supplementary Material (Data Sheets 1–3).

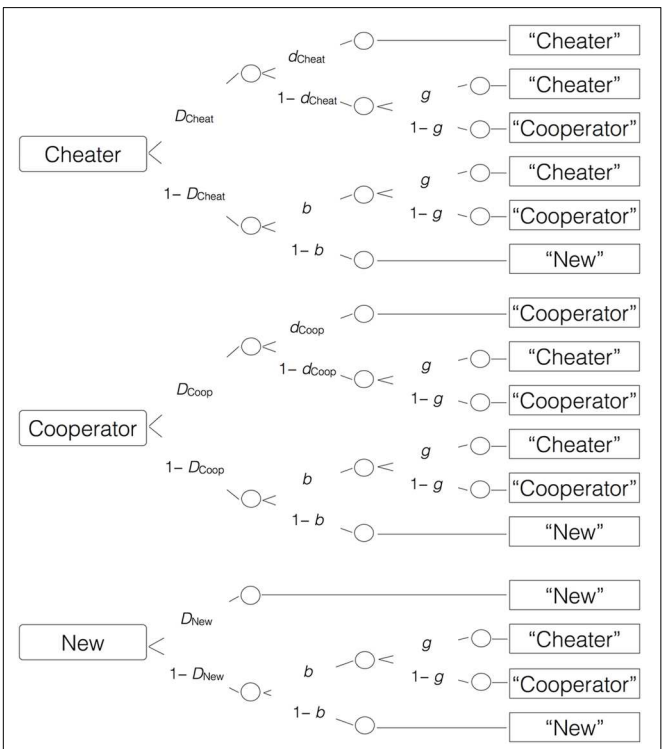


FIGURE 2 | The multinomial source memory model adapted from Bayen et al. (1996). Rounded rectangles on the left represent the items presented in the source memory test (cheater, cooperator, or new faces). The letters along the branches represent the probabilities with which certain memory states occur (*D*: probability to correctly recognize a face as old or new; *d*: conditional probability to correctly remember that the person was a cheater or a cooperator; *g*: conditional probability to guess that the person was a cheater; *b*: conditional probability to guess that a face was old). Rectangles on the right represent the participants' responses in the memory test.

Results

Game Investments

Game investments were analyzed with a repeated measures MANOVA with facial trustworthiness (trustworthy-looking vs. untrustworthy-looking) as independent variable. Participants only interacted once with each partner and thus had no chance to anticipate the behavior of the partners before they decided whether to invest or not. Therefore, only the partners' facial trustworthiness, but not their behavior could influence the investments. As expected, participants invested more money when playing with trustworthy-looking partners than when playing with untrustworthy-looking partners, $F(1,111) = 136.83$, $p < 0.001$, $\eta_p^2 = 0.55$ (see left panel of **Figure 3**).

Likability Ratings

Likability ratings were analyzed with a 2 × 2 MANOVA with facial trustworthiness (trustworthy-looking vs. untrustworthy-looking) and partner behavior (cheating vs. cooperation) as independent variables. Trustworthy-looking faces were more likable than untrustworthy-looking faces, $F(1,111) = 410.29$, $p < 0.001$, $\eta_p^2 = 0.79$. Cooperators received higher likability ratings than cheaters, $F(1,111) = 12.94$, $p < 0.001$, $\eta_p^2 = 0.10$. There was no interaction between facial trustworthiness and behavior, $F(1,111) = 1.75$, $p = 0.189$, $\eta_p^2 = 0.01$ (see left panel of **Figure 4**).

Old–New Recognition

Old–new recognition in terms of P_r (the sensitivity measure of the two-high-threshold model of old–new recognition, often referred to as corrected hit rate and given by hit rate minus false alarm rate; Snodgrass and Corwin, 1988) is shown in the left panel of **Figure 5**. A 2 × 2 MANOVA was performed with facial trustworthiness (trustworthy-looking vs. untrustworthy-looking) and partner behavior (cheating vs. cooperation) as independent variables. There was no main effect of facial trustworthiness on face recognition, $F(1,111) = 0.52$, $p = 0.472$, $\eta_p^2 < 0.01$, no main effect of partner behavior, $F(1,111) = 1.11$, $p = 0.294$, $\eta_p^2 = 0.01$, and no interaction between facial trustworthiness and behavior, $F(1,111) = 0.90$, $p = 0.346$, $\eta_p^2 < 0.01$.

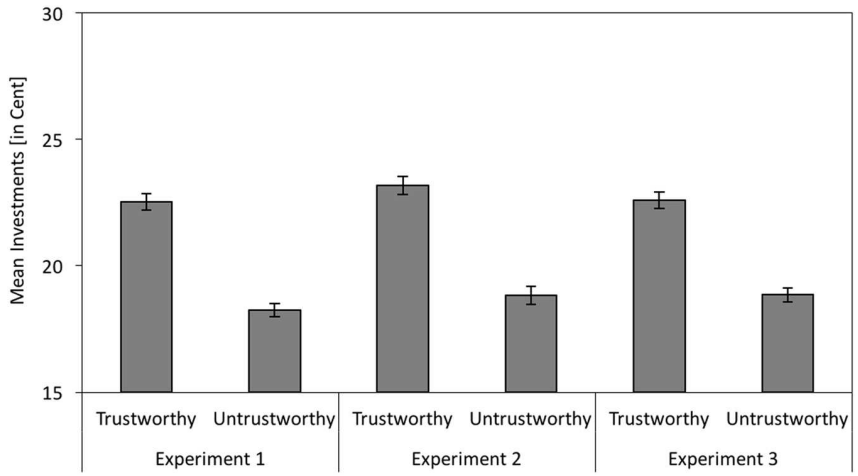


FIGURE 3 | Participants' mean investments in the social interaction game as a function of facial trustworthiness (trustworthy vs. untrustworthy) in Experiment 1 (without cognitive load) and in Experiments 2 and 3 (with cognitive load). The error bars represent the standard errors.

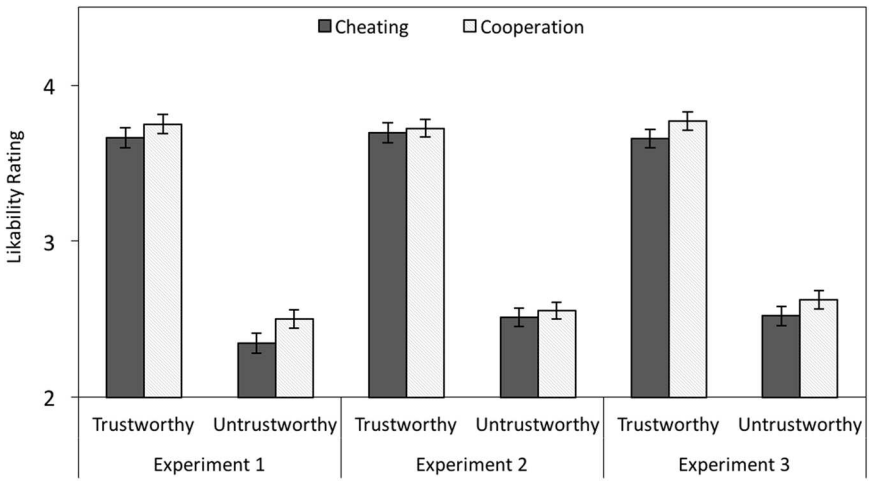


FIGURE 4 | Mean test-phase likability ratings (on a scale ranging from 1 to 6) as a function of facial trustworthiness (trustworthy vs. untrustworthy) and behavior (cheating vs. cooperation) in Experiment 1 (without cognitive load) and in Experiments 2 and 3 (with cognitive load). The error bars represent the standard errors.

Source Guessing and Source Memory

To disentangle source guessing and memory, the multinomial source monitoring model mentioned above (Bayen et al., 1996) was used. For the present study, we needed two sets of the trees displayed in **Figure 2**, one for trustworthy faces and one for untrustworthy faces. To obtain an identifiable base model, we assumed that old–new recognition does not differ as a function of partner behavior (as evidenced by the analysis of old–new recognition reported above), and does not differ between old and new faces ($D_{\text{Cheat}} = D_{\text{Coop}} = D_{\text{New}}$), which is commonly assumed when using the two high threshold model (Snodgrass and Corwin, 1988; Bayen et al., 1996). This base model fit the data well, $G^2(2) = 1.84$, $p = 0.398$.

First, we analyzed whether participants would show an expectancy-congruent guessing bias. When the behavior of a recognized person is not remembered, participants have to guess whether the face was associated with cheating or cooperation. In previous studies (Bell et al., 2012b), participants guessed that trustworthy-looking persons were cooperators and that untrustworthy-looking persons were cheaters. That pattern was replicated here. If source memory was not available at test, participants showed a strong bias toward guessing that trustworthy-looking faces were previously associated with cooperation and that untrustworthy-looking faces were previously associated with cheating, $\Delta G^2(1) = 43.01$, $p < 0.001$, $w = 0.07$ (see left panel of **Figure 6**).

The left panel of **Figure 7** displays the estimates for source memory parameter d representing the conditional probability of remembering the behaviors of cheaters and cooperators given that their faces were recognized as old. Source memory was

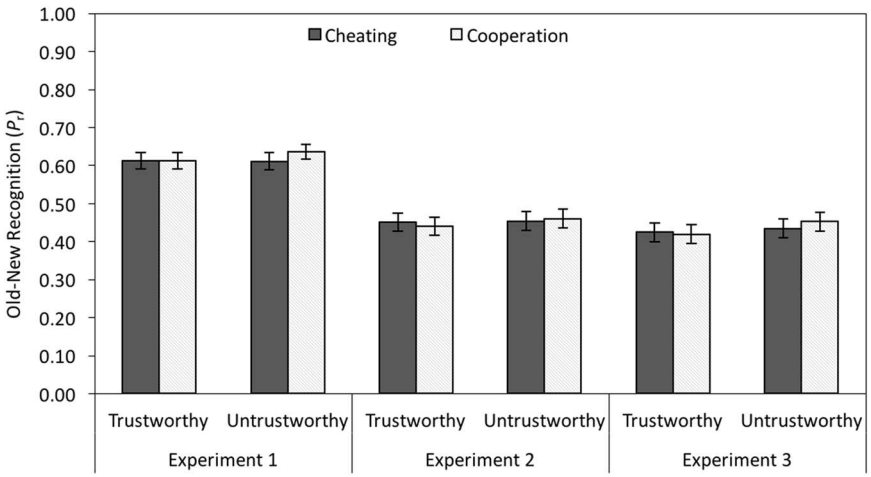


FIGURE 5 | Old–new recognition in terms of P_r (corrected hit rates) as a function of facial trustworthiness (trustworthy vs. untrustworthy) and partner behavior (cheating vs. cooperation) in Experiment 1 (without cognitive load) and in Experiments 2 and 3 (with cognitive load). The error bars represent the standard errors.

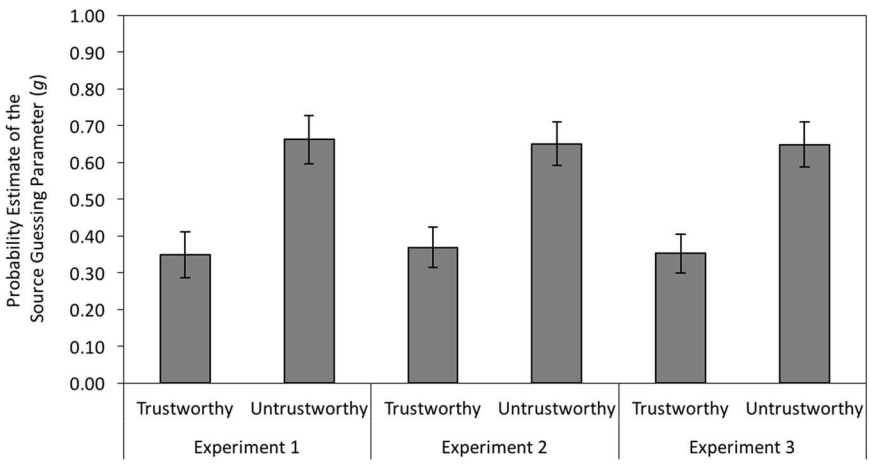


FIGURE 6 | Estimates of the guessing parameter *g* representing the probability to guess that a person was a cheater rather than a cooperator as a function of facial trustworthiness (trustworthy vs. untrustworthy) in Experiment 1 (without cognitive load) and in Experiments 2 and 3 (with cognitive load). The error bars represent the 95% confidence intervals.

better for cheaters than for cooperators when the faces looked trustworthy, $\Delta G^2(1) = 4.82$, $p = 0.028$, $w = 0.02$, but there was no corresponding memory advantage for cooperators over cheaters when the faces looked untrustworthy, $\Delta G^2(1) = 0.14$, $p = 0.704$, $w < 0.01$. Thus, we replicated the finding of an asymmetrical expectancy-violation effect (Suzuki and Suga, 2010; Bell et al., 2012b).

Discussion

In Experiment 1, as in previous studies (van 't Wout and Sanfey, 2008; Bell et al., 2012b, 2013), participants invested more money into the sequential Prisoner's Dilemma Game (trusted their partners more) when the partners looked trustworthy than when they looked untrustworthy. In the memory test, old–new recognition was not affected by facial trustworthiness and partner behavior, consistent with a large number of previous studies showing that a person's behavior has no effect on old–new face recognition (e.g., Barclay and Lalumière, 2006; Mehl and Buchner, 2008; Buchner et al., 2009; Kroneisen and Bell, 2013). There are some reports suggesting that old–new recognition is better for untrustworthy-looking than for trustworthy-looking persons (Rule et al., 2012; Bell et al., 2013; Mattarozzi et al., 2015), but this finding was not reliably obtained across experiments (Bell et al., 2012b), and was not replicated here. Consistent with several other studies (Nash et al., 2010; Bell et al., 2012b; Cassidy et al., 2012), participants demonstrated a bias toward guessing that trustworthy-looking persons were cooperators and untrustworthy-looking persons were cheaters. Moreover, and in line with previous studies (Suzuki and Suga, 2010; Bell et al., 2012b), an asymmetric source memory advantage

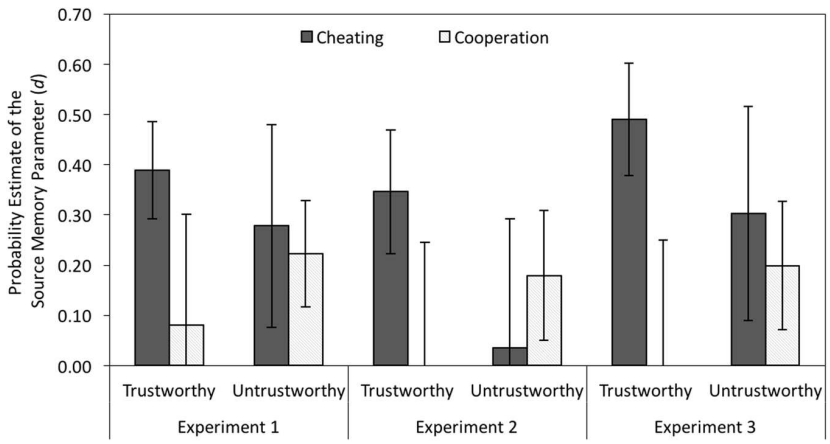


FIGURE 7 | Estimates of the source memory parameter *d* as a function of the partners' facial trustworthiness (trustworthy vs. untrustworthy) and the partners' behavior (cheating vs. cooperation) in Experiment 1 (without cognitive load) and in Experiments 2 and 3 (with cognitive load). The error bars represent the 95% confidence intervals.

for appearance-incongruent negative information was found: Participants had better source memory for trustworthy-looking cheaters than for trustworthy-looking cooperators.

EXPERIMENT 2

Experiment 2 served to test whether a different pattern of results would be obtained under cognitive load. To impose cognitive load, a continuous choice reaction time (CRT) task with auditory stimuli was used as secondary task. This is a well established method to impose cognitive load (Naveh-Benjamin et al., 2003; Kroneisen et al., 2014), and has the advantage that it involves non-verbal stimuli and responses that do not directly interfere with the sequential Prisoner's Dilemma Game. Participants had to classify three randomly varying tones by pressing three buttons on a response box. The tones were continuously presented to guarantee a steady burden on cognitive resources. The main question was whether the expectancy-violation effect on source memory would disappear under conditions of reduced cognitive resources.

Method

Participants

One hundred and nine students (67 of whom were female) with a mean age of 24 ($SD = 5$) participated in Experiment 2. Participants in Experiment 2 did not differ from those in Experiment 1 in terms of age, gender, and justice sensitivity (**Table 1**). All participants gave written informed consent.

Materials, Procedure, and Design

Experiment 2 was identical to Experiment 1 except that participants were required to perform a secondary CRT task during the sequential Prisoner's Dilemma Game. The task was to continuously classify three piano tones (C1, F3, and B6) by pressing a black left, gray middle, or white right button on a response box, respectively. Each tone was repeated once every second until participants made a CRT response by pressing a CRT button. Participants received no reminder of the CRT task and no explicit warning when they failed to respond to the CRT stimuli (but the repeated presentation of the same tone can be seen as an implicit warning). Before the start of the sequential Prisoner's Dilemma Game, participants received a training of the CRT task. During this training, participants received immediate feedback about their responses ("correct" in green font color or "false" or "miss" in red font color). This training continued until participants had 20 correct responses in a row.

Given that participants were not pressured to perform the secondary CRT task, it was necessary to exclude participants who did not respond to the CRT stimuli properly. As an inclusion criterion, we required a minimum of one response per trial in the Prisoner's Dilemma Game on average. Based on this criterion, datasets of 13 participants were excluded from analyses because of too few CRT responses. With the remaining sample consisting of 96 participants, it was possible to detect an effect of size $w = 0.04$ for the comparison of source memory between cheaters and cooperators with a statistical power $(1 - \beta)$ of 0.94.

Results

Game Investments

As in Experiment 1, participants invested more when playing with trustworthy-looking partners than when playing with untrustworthy-looking partners, $F(1,95) = 160.64$, $p < 0.001$, $\eta_p^2 = 0.63$ (see middle panel of **Figure 3**).

Likability Ratings

There was a main effect of facial trustworthiness on likability, $F(1,95) = 433.80$, $p < 0.001$, $\eta_p^2 = 0.82$. The effect of partner behavior was not significant, $F(1,95) = 1.13$, $p = 0.290$, $\eta_p^2 = 0.01$. There was no interaction between facial trustworthiness and behavior, $F(1,95) = 0.07$, $p = 0.794$, $\eta_p^2 < 0.01$ (see middle panel of **Figure 4**).

Old–New Recognition

Old–new recognition was lower than in Experiment 1, but the same pattern of results was obtained (see middle panel of **Figure 5**). There was neither a main effect of facial trustworthiness, $F(1,95) = 0.34$, $p = 0.563$, $\eta_p^2 < 0.01$, nor a main effect of partner behavior, $F(1,95) = 0.02$, $p = 0.897$, $\eta_p^2 < 0.01$. The two-way interaction was not significant, $F(1,95) = 0.34$, $p = 0.562$, $\eta_p^2 < 0.01$.

Source Guessing and Source Memory

The base model fit the data well, $G^2(2) = 0.32$, $p = 0.852$. As in Experiment 1, participants were more likely to guess that untrustworthy-looking faces were associated with cheating than that trustworthy-looking faces were associated with cheating, $\Delta G^2(1) = 48.32$, $p < 0.001$, $w = 0.08$ (see middle panel of **Figure 6**).

Again, source memory was better for cheating than for cooperation when the faces looked trustworthy, $\Delta G^2(1) = 5.22$, $p = 0.022$, $w = 0.03$, and source memory did not differ between cheating and cooperation when the faces looked untrustworthy, $\Delta G^2(1) = 0.67$, $p = 0.414$, $w < 0.01$ (see middle panel of **Figure 7**).

Performance in the Continuous Reaction Time Task

The description of the results is incomplete without an analysis of the performance in the CRT task because it is important to test whether or not the enhanced memory for appearance-incongruent cheating is due to a performance trade-off between the encoding of the faces and the CRT task. Therefore, we performed two 2 × 2 MANOVAs with the partner trustworthiness (trustworthy-looking vs. untrustworthy-looking) and partner behavior (cheating vs. cooperation) as independent variables and the proportion of correct responses and the response times (including only correct responses that occurred after > 100 ms) in the CRT task as dependent variables (**Table 2**). Proportion correct did not differ as a function of facial trustworthiness, $F(1,95) = 2.43$, $p = 0.122$, $\eta_p^2 = 0.02$. However, CRT performance was less accurate in the cheater condition in comparison to the cooperator condition, $F(1,95) = 5.76$, $p = 0.018$, $\eta_p^2 = 0.06$. There was no interaction between facial trustworthiness and partner behavior, $F(1,95) = 0.14$, $p = 0.704$, $\eta_p^2 < 0.01$. Response times showed a similar pattern. Response time did not differ as a function of facial trustworthiness, $F(1,95) = 0.31$, $p = 0.578$, $\eta_p^2 < 0.01$. Responses were slower in the cheater condition in comparison to the cooperator condition, $F(1,95) = 5.09$, $p = 0.026$, $\eta_p^2 = 0.05$. However, there was no interaction between facial trustworthiness and partner behavior, $F(1,95) = 0.15$, $p = 0.697$, $\eta_p^2 < 0.01$. Given that this attentional disruption did not translate into better memory for cheaters (as shown by the analyses above), this result does not seem to reflect a reallocation of cognitive resources to the cheater faces and, therefore, does not seem to reflect a performance trade-off between the memory task and the CRT task. It seems possible to speculate that experiencing cheating may result in a negative emotional response that may distract from the secondary task, but does not seem to cause a direct memory enhancement.

TABLE 2 | Mean proportion correct and response times in milliseconds in the CRT task as a function of the partners' facial trustworthiness (trustworthy vs. untrustworthy) and the partners' behavior (cheating vs. cooperation) in Experiments 2 and 3.

	Cheating		Cooperation	
	M	*SE*	*M*	*SE*
Experiment 2				
Proportion correct				
Trustworthy Faces	0.90	0.01	0.92	0.01
Untrustworthy Faces	0.91	0.01	0.92	0.01
Response time				
Trustworthy Faces	2,252	79	2,186	86
Untrustworthy Faces	2,240	99	2,149	86
Experiment 3				
Proportion correct				
Trustworthy Faces	0.88	0.01	0.90	0.01
Untrustworthy Faces	0.89	0.01	0.90	0.01
Response time				
Trustworthy Faces	788	13	762	13
Untrustworthy Faces	787	12	760	13

Discussion

Even though participants had to perform a secondary CRT task, the results were almost identical to those of Experiment 1. Most importantly, participants showed evidence of an appearance-congruent guessing bias and of an asymmetrical expectancy-violation effect on source memory. We conclude from these findings that the enhanced memory for expectancy-incongruent information is obtained even under conditions of cognitive load, which suggests that the encoding of this information occurs automatically and does not rely on demanding elaborative processes.

It seemed important to address the possible concern that the CRT task may simply not have been demanding enough to interfere with the primary task. In Experiment 2, participants were required to perform the secondary CRT task concurrently to the Prisoner's Dilemma Game, but no time pressure was imposed. Therefore, it may have been possible to attend to both the CRT task and the Prisoner's Dilemma Game by delaying responses in the CRT task. In Experiment 3, we therefore required participants to respond to each tone within a time interval of 2 s (which is a typical time interval in CRT studies, see Kroneisen et al., 2014).

EXPERIMENT 3

Experiment 3 was identical to Experiment 2 with the exception that the CRT task was modified to increase the continuous demands on cognitive resources.

Method

Participants

One hundred three students (69 of whom were female) with a mean age of 22 ($SD = 5$) participated in Experiment 3. The

sample was similar to those in Experiments 1 and 2 (**Table 1**). All participants gave written informed consent.

Materials, Procedure, and Design

Experiment 3 was identical to Experiment 2 with the exception that the CRT task required participants to respond to each tone within 2 s, after which the next tone was presented. If participants failed to respond to a tone during a trial of the sequential Prisoner's Dilemma Game, they received a warning after the trial that reminded them of the CRT task. In contrast to Experiment 2—in which the sequential Prisoner's Dilemma Game was self-paced—the next round of the game was automatically initiated 10 s after the summary of the interaction had been displayed. Justice sensitivity was not assessed.

The data of two outliers were excluded from the analyses because these participants produced >20% CRT misses on average. The remaining sample responded to 98% of the CRT stimuli on average. With a remaining sample of 101 participants, it was possible to detect an effect of size $w = 0.04$ for the comparison between source memory for cheaters and cooperators with a statistical power $(1 - \beta)$ of 0.95.

Results

Game Investments

As in Experiments 1 and 2, participants invested more when playing with trustworthy-looking partners than when playing with untrustworthy-looking partners, $F(1,100) = 157.95$, $p < 0.001$, $\eta_p^2 = 0.61$ (see right panel of **Figure 3**).

Likability Ratings

There was a main effect of facial trustworthiness on likability with higher likability ratings for trustworthy-looking compared to untrustworthy-looking partners, $F(1,100) = 504.95$, $p < 0.001$, $\eta_p^2 = 0.83$. Cheaters were judged to be less likable than cooperators, $F(1,100) = 15.08$, $p < 0.001$, $\eta_p^2 = 0.13$. The interaction between facial trustworthiness and behavior was not significant, $F(1,100) = 0.05$, $p = 0.822$, $\eta_p^2 < 0.01$ (see right panel of **Figure 4**).

Old–New Recognition

There was neither a main effect of facial trustworthiness on old–new recognition, $F(1,100) = 1.49$, $p = 0.225$, $\eta_p^2 = 0.01$, nor a main effect of partner behavior, $F(1,100) = 0.21$, $p = 0.651$, $\eta_p^2 < 0.01$. The two-way interaction was also not significant, $F(1,100) = 0.57$, $p = 0.452$, $\eta_p^2 < 0.01$ (see right panel of **Figure 5**).

Source Guessing and Source Memory

The base model fit the data well, $G^2(2) = 0.87$, $p = 0.647$. As in Experiments 1 and 2, participants were significantly more likely to guess that untrustworthy-looking faces were associated with cheating than that trustworthy-looking faces were associated with cheating, $\Delta G^2(1) = 55.78$, $p < 0.001$, $w = 0.08$ (see right panel of **Figure 6**).

As in the previous experiments, there was a source memory advantage for cheaters over cooperators when the faces looked trustworthy, $\Delta G^2(1) = 12.60$, $p < 0.001$, $w = 0.04$, but source memory did not differ between cheaters and cooperators when the faces looked untrustworthy, $\Delta G^2(1) = 0.42$, $p = 0.519$, $w < 0.01$ (see right panel of **Figure 7**).

Performance in the Continuous Reaction Time Task

As in Experiment 2 we performed analyses of the proportion of correct responses and response times (including only correct responses that occurred after >100 ms) in the CRT task. CRT responses were faster than they were in Experiment 2, but the same pattern of results was observed (**Table 2**). Proportion correct did not differ as a function of facial trustworthiness, $F(1,100) = 0.42$, $p = 0.520$, $\eta_p^2 < 0.01$. CRT performance was less accurate in the cheater condition in comparison to the cooperator condition, $F(1,100) = 21.82$, $p < 0.001$, $\eta_p^2 = 0.18$. There was no interaction between facial trustworthiness and partner behavior, $F(1,100) = 0.55$, $p = 0.460$, $\eta_p^2 < 0.01$. Response times showed a similar pattern. Response time did not differ as a function of facial trustworthiness, $F(1,100) = 0.09$, $p = 0.764$, $\eta_p^2 < 0.01$. However, responses were slower in the cheater condition in comparison to the cooperator condition, $F(1,100) = 33.29$, $p < 0.001$, $\eta_p^2 = 0.25$. There was no interaction between facial trustworthiness and partner behavior, $F(1,100) = 0.04$, $p = 0.845$, $\eta_p^2 < 0.01$. Again, the previous analyses suggest that this attentional disruption is not associated with enhanced encoding of the cheater faces.

Discussion

Even though participants were pressured to make faster responses in the CRT task, the same pattern of results was obtained as in Experiments 1 and 2. Most importantly, we obtained evidence in favor of an expectancy-congruent guessing bias and of an asymmetric expectancy-violation effect. Therefore, it seems possible to conclude that the encoding of expectancy-incongruent information works well even under conditions of high cognitive load, presumably because it occurs automatically. At a descriptive level, the results of all three experiments are strikingly similar with the only exception that old–new recognition seems to be somewhat decreased in Experiments 2 and 3 in comparison to Experiment 1.

Given that the CRT task did not seem to have any substantial effect on source memory (or any other variable except face recognition), it may be tempting to conclude from these findings that the CRT task was simply not demanding enough. However, concluding from a non-significant finding that the cognitive load manipulation was not strong enough is problematic because this type of circular reasoning renders the prediction that cognitive load affects cooperation and memory unfalsifiable. To escape this problem, we performed a validation study to test whether the secondary task does indeed disrupt cognitively demanding working-memory processes (as intended).

EXPERIMENT 4

Experiment 4 served to validate the CRT task by testing whether it does indeed have the capacity to disrupt cognitively demanding processes. We used both a verbal memory task and a spatial memory task to test whether the CRT task interferes generally with cognitive processing and does not only selectively affect the

processing of a specific type of information (Lange, 2005; Vachon et al., in press).

Method

Participants

Forty students (27 of whom were female) with a mean age of 24 (SD = 4) participated in Experiment 4. Participants were consecutively assigned to either the cognitive load group or the control group (i.e., Participant 1 was assigned to the cognitive load condition, Participant 2 was assigned to the control condition, and so on). All participants gave written informed consent.

Materials, Procedure, and Design

Participants performed a verbal working memory task and a spatial working memory task. Task order was counterbalanced between groups (cognitive load vs. control).

In the verbal working memory task, participants were required to remember sequences with varying sequence lengths of four to nine items. The items were randomly drawn from the set {1, 2, . . . 9}. Each trial started with a visual warning that participants were required to remember the digits. The digits were presented one after another in 24 pt Arial font at the center of a computer screen for 800 ms with a 200 ms inter-stimulus interval. After a retention interval of 2 s, a number pad with the previously presented digits was shown, and participants were required to select the numbers in the correct (forward) order, using the computer mouse. Selected digits were grayed out, and could not be selected again. After all digits were selected, the number pad disappeared, and a continue button was shown. Upon clicking this button, the next trial started. The task started with a sequence length of four digits. Digit length gradually increased during the task. Participants completed three trials of each sequence length.

The spatial working memory task was identical to the verbal working memory task except that participants were required to remember the spatial locations of four to nine black dots instead of four to nine digits. The locations of the dots were not aligned (but instead randomly distributed across the screen) to make a verbal coding strategy extremely difficult. In each trial, the spatial positions were randomly drawn from a set of nine different spatial positions. The dots appeared one after another at their designated positions (800 ms on, 200 ms off). After a retention interval of 2 s, the previously presented dots were presented again at their corresponding spatial locations. The participants' task was to select the spatial locations of the dots in the order of their appearance. Selected locations were grayed out, and could not be selected again.

The working memory tasks were either completed alongside the secondary CRT task (in the cognitive load condition) or without the secondary CRT task (in the control condition). The CRT task was identical to the one used in Experiment 3. Participants were reminded of the tone classification task before each trial. Tones were presented only during visual item presentation and the retention interval of the working memory task, but not during recall. If participants did not give a response to all CRT tones, they received a warning when the recall of the items was completed.

The design was a mixed 2 × 2 design with working memory task (verbal vs. spatial) as a within-subject variable and cognitive load (cognitive load vs. control) as a between-subjects variable. The dependent variable was working memory performance according to a strict scoring criterion (only items remembered in their correct serial position were scored as correct). Given $\alpha = 0.05$, a total sample size of $N = 40$ participants, and an assumed correlation between the levels of the within-subject variable of $\rho = 0.50$, an effect of size $f = 0.50$ could be detected for the cognitive load variable with a statistical power $(1 - \beta)$ of 0.95.

Results

A 2 × 2 MANOVA with cognitive load (cognitive load vs. control) and working memory task (verbal vs. spatial) as independent variables yielded a main effect of cognitive load, $F(1,38) = 20.60$, $p < 0.001$, $\eta_p^2 = 0.35$, and of task, $F(1,38) = 70.34$, $p < 0.001$, $\eta_p^2 = 0.65$, but no interaction between cognitive load and task, $F(1,38) = 1.75$, $p = 0.193$, $\eta_p^2 = 0.04$. Cognitive load significantly decreased memory performance both in the verbal, $t(38) = 3.68$, $p = 0.001$, $\eta_p^2 = 0.26$, and in the spatial task, $t(38) = 4.05$, $p < 0.001$, $\eta_p^2 = 0.30$ (**Figure 8**). Raw data are reported in the Online Supplementary Material (Data Sheet 4).

Discussion

Experiment 4 serves as a validation study to confirm that the CRT task interferes with cognitively demanding processes. In line with our expectations, the CRT task disrupted performance in a verbal working memory task as well as in a spatial working memory task, suggesting that it does not only interfere with a specific type of information processing, but instead leads to a general decrease of cognitive resources. This rules out the possibility that the CRT

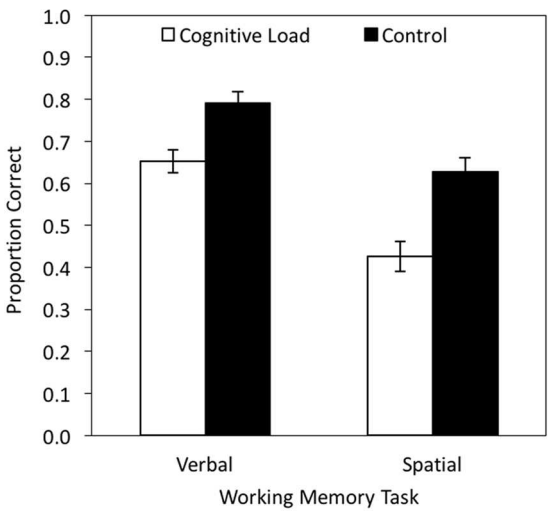


FIGURE 8 | Mean working memory performance in proportion correct as a function of working memory task (verbal vs. spatial) and cognitive load (cognitive load vs. control). The error bars represent the standard errors.

task was not demanding enough to disrupt cognitive processing, which facilitates the interpretation of the findings obtained in Experiments 1–3.

GENERAL DISCUSSION

Previous research suggests that expectations about other people's trustworthiness are formed quickly and automatically on the basis of physical appearance (Todorov et al., 2009, 2015). Trustworthiness judgments in particular are strongly affected by facial cues (Todorov, 2008). The assumption that facial cues have a strong effect on trust and social expectations (van 't Wout and Sanfey, 2008) is further confirmed by the present results. Specifically, participants invested more into the sequential Prisoner's Dilemma Game when the partners looked trustworthy than when the partners looked untrustworthy. Given that investing into the game only payed off when the partner reciprocated, this result suggests that trustworthy-looking partners were expected to cooperate more than untrustworthy-looking partners. Noticeably, this pattern of results was obtained without and with cognitive load, which confirms previous findings suggesting that the perception of facial trustworthiness is an automatic process that does not depend on the availability of cognitive resources (Bonnefon et al., 2013).

Given that appearance-based judgments about a person are often invalid (Todorov et al., 2015), it is important to update facial trustworthiness judgments with behavioral information (Rezlescu et al., 2012). It may be especially important to remember expectancy-incongruent behaviors to be able to correct a false first impression about another person. Consistent with previous studies (Suzuki and Suga, 2010; Volstorf et al., 2011; Bell et al., 2012b), source memory was better for the appearance-incongruent cheating of a trustworthy-looking person in comparison to the appearance-congruent cooperation of a trustworthy-looking person. Noticeably, memory for appearance-congruent cooperation was poor. This confirms the predictions of the schema-copy-plus-tag model (Graesser and Nakamura, 1982), which states that discriminability of schema-consistent information is poor because it will be produced at test regardless of whether it was presented at encoding or not. Schema-atypical information is more distinct, and, therefore, associated with better memory discriminability.

Memory was selectively enhanced for cheating that violated a positive expectation about a trustworthy-looking partner, but there was no similar memory advantage for cooperators over cheaters when the faces looked untrustworthy. This asymmetry was also found in previous memory experiments (Suzuki and Suga, 2010; Bell and Buchner, 2012), and it fits with a study on investments in repeated game interactions showing that participants tend to adjust their own behavior more strongly in response to a partner's defection than in response to a partner's cooperation (Chang et al., 2010). This asymmetric memory advantage for appearance-incongruent cheating over appearance-incongruent cooperation may be particularly pronounced in the present study because only female stimulus faces were used. It is known that female faces tend to elicit positive social expectations (Kroneisen and Bell, 2013), which means that norm-violating behaviors of female partners may represent particularly strong expectancy violations (Bell et al., 2015).

Two explanations for the memory advantage for appearance-incongruent cheating were tested. According to the first account, information that does not fit into existing schemas receives more elaborative processing, which depends on the mobilization and availability of additional cognitive resources. This enhanced elaboration results in a more vivid and detailed recollection of the expectancy-incongruent information. According to the second account, schema-atypical information is retained in form of unelaborated tags. This resource-efficient encoding strategy has the advantage that unexpected information can be encoded and retained in memory even under conditions of high cognitive load. The present results support the latter view. The source memory advantage for appearance-incongruent cheating was not affected by the presence or absence of cognitive load at encoding. A similar memory advantage for appearance-incongruent cheating was obtained in all three experiments, regardless of whether participants had to perform a demanding secondary task at encoding or not. The experiments were reported separately because they were run at different times. However, when the source memory data of all experiments were combined in a single supplementary cross-experimental analysis, the conclusion that source memory was not affected by cognitive load was supported. The base model still fit the data well, $G^2(6) = 3.02$, $p = 0.807$. Source memory did not differ among experiments, $\Delta G^2(8) = 11.95$, $p = 0.154$, $w = 0.02$, which suggests that the pattern of results was not affected by the secondary task in Experiments 2 and 3.

This pattern of findings confirms the predictions of the schema-copy-plus-tag model (Graesser and Nakamura, 1982), according to which schema-violating information is retained in the form of simple tags that require only minimal elaboration, and can therefore be encoded and retained even under conditions of high cognitive load. Consistent with this interpretation, it has been previously shown that the source memory advantage for faces of cheaters is not due to a vivid recollection of the cheating episode, but rather due to emotional tagging in the sense of a rough classification of the partner as a "cheater" (Bell et al., 2012a). The encoding and retrieval of simple emotional tags may be less cognitively demanding and, therefore, less affected by a reduction in cognitive resources than other types of context memory (Rahhal et al., 2002).

This interpretation fits well with Todorov and Uleman's (2003) assumption that reading about or observing the behavior of another person leads people to draw inferences about the other person's traits (e.g., dishonest or honest) that then become linked to the other person's face. Importantly, these trait representations are assumed to include only a summary judgment about the other person's behavior, and to be comparatively unelaborated and robust (Carlston and Skowronski, 1994; Todorov and Uleman, 2002). In the study of Todorov and Uleman (2003), participants saw faces with behavior descriptions that implied character traits.

The binding between faces and traits was revealed by an enhanced false recognition of the trait labels in an implicit memory test. The most interesting finding in the present context is that the implicit memory for the association between a face and a trait was not affected by a secondary task at encoding (rehearsing 6-digit numbers), which suggests that the process of binding traits to faces is an automatic process. The present study shows parallel findings in a different paradigm where traits are directly inferred from experiences in a social-dilemma game, and memory is tested in an explicit source memory test.

Remembering appearance-congruent cooperation and cheating enables participants to update their impressions about other people, which could have beneficial effects on future social decision making. For instance, when we encounter a trustworthy-looking person, but learn subsequently that this person is not to be trusted, memory for the appearance-incongruent cheating may help to avoid being fooled by the trustworthy appearance of this person again. Obviously, this discussion implies that the memory for the partners' previous behaviors is used to inform social decision making. Previous results using repeated social-dilemma games suggest that people continue to rely on facial trustworthiness over the course of the game (in line with the persistent effect of facial trustworthiness on source guessing in the present experiment), but also succeed in adjusting their own decisions to the individual partners' previous trustworthy or untrustworthy behaviors (Chang et al., 2010; Rezlescu et al., 2012). Murty et al. (2016) directly examined the relationship between memory and economic decision making, and found that source memory (in contrast to item memory) had a beneficial effect on the participants' choices in social and non-social decision making tasks. Therefore, it seems plausible to assume that source memory for appearance-incongruent behaviors can have direct beneficial effects on social decision making.

CONCLUSION

In sum, source memory for cheaters and cooperators was highly similar across experiments, regardless of whether cognitive load was induced at encoding (Experiments 2 and 3) or not (Experiment 1). These results are compatible with the general idea that cognitive mechanisms underlying social cooperation operate highly automatically so that they remain unaffected by cognitive load. Specifically, it seems possible to encode and retain information about a person's expectancy-incongruent behavior even under conditions of high cognitive load. Remembering this type of behavior seems particularly important for the decision making process because it helps to correct maladaptive behavior tendencies. For example, it seems particularly important to remember that a trustworthy-looking person is in fact not to be trusted to avoid being fooled by the trustworthy appearance of this person in the future. Being able to remember appearance-incongruent behaviors even under conditions of cognitive load may be beneficial in that it allows people to sustain successful reciprocal cooperation even under the distracting and stressful conditions that are characteristic of everyday life.

AUTHOR CONTRIBUTIONS

LM, RB, and AB conceived and designed the experiments, supervised data collection, analyzed the data, and wrote the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01312

REFERENCES

Barclay, P., and Lalumière, M. (2006). Do people differentially remember cheaters? *Hum. Nat.* 17, 98–113. doi: 10.1007/s12110-006-1022-y

Bayen, U. J., Murnane, K., and Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 197–215.

Bell, R., and Buchner, A. (2012). How adaptive is memory for cheaters? *Curr. Dir. Psychol. Sci.* 21, 403–408. doi: 10.1177/0963721412458525

Bell, R., Buchner, A., Erdfelder, E., Giang, T., Schain, C., and Riether, N. (2012a). How specific is source memory for faces of cheaters? Evidence for categorical emotional tagging. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 457–472. doi: 10.1037/a0026017

Bell, R., Buchner, A., Kroneisen, M., and Giang, T. (2012b). On the flexibility of social source memory: a test of the emotional incongruity hypothesis. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 1512–1529. doi: 10.1037/a0028219

Bell, R., Giang, T., Mund, I., and Buchner, A. (2013). Memory for reputational trait information: is social–emotional information processing less flexible in old age? *Psychol. Aging* 28, 984–995. doi: 10.1037/a0034266

Bell, R., Mieth, L., and Buchner, A. (2015). Appearance-based first impressions and person memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 456–472. doi: 10.1037/xlm0000034

Bonnefon, J.-F., Hopfensitz, A., and De Neys, W. (2013). The modular nature of trustworthiness detection. *J. Exp. Psychol. Gen.* 142, 143–150. doi: 10.1037/a0028930

Bröder, A., and Meiser, T. (2007). Measuring source memory. *Z. Psychol. J. Psychol.* 215, 52–60.

Buchner, A., Bell, R., Mehl, B., and Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evol. Hum. Behav.* 30, 212–224. doi: 10.1016/j.evolhumbehav.2009.01.004

Carlston, D., and Skowronski, J. (1994). Savings in the relearning of trait information as evidence for spontaneous inference generation. *J. Pers. Soc. Psychol.* 66, 840–856. doi: 10.1037/0022-3514.66.5.840

Cassidy, B. S., Zebrowitz, L. A., and Gutchess, A. H. (2012). Appearance-based inferences bias source memory. *Mem. Cogn.* 40, 1214–1224. doi: 10.3758/s13421-012-0233-1

Chang, L. J., Doll, B., van 't Wout, M., Frank, M., and Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cogn. Psychol.* 61, 87–105. doi: 10.1016/j.cogpsych.2010.03.001

Clark, K., and Sefton, M. (2001). The sequential prisoner's dilemma: evidence on reciprocation. *Econ. J.* 111, 51–68. doi: 10.1111/1468-0297.00588

Delton, A., Krasnow, M., Cosmides, L., and Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13335–13340. doi: 10.1073/pnas.1102131108

Ehrenberg, K., and Klauer, K. C. (2005). Flexible use of source information: processing components of the inconsistency effect in person memory. *J. Exp. Soc. Psychol.* 41, 369–387. doi: 10.1016/j.jesp.2004.08.001

Engell, A. D., Haxby, J. V., and Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* 19, 1508–1519. doi: 10.1162/jocn.2007.19.9.1508

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., and Nadarevic, L. (2009). Multinomial processing tree models. *Z. Psychol. J. Psychol.* 217, 108–124.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Graesser, A. C., and Nakamura, G. V. (1982). "The impact of a schema on comprehension and memory," in *The Psychology of Learning and Motivation*, ed. G. H. Bower (London: Academic Press), 59–109.

Kessler, J., and Meier, S. (2014). Learning from (failed) replications: cognitive load manipulations and charitable giving. *J. Econ. Behav. Organ.* 102, 10–13. doi: 10.1016/j.jebo.2014.02.005

Klein, S., Cosmides, L., Tooby, J., and Chance, S. (2002). Decisions and the evolution of memory: multiple systems, multiple functions. *Psychol. Rev.* 109, 306–329. doi: 10.1037/0033-295X.109.2.306

Kroneisen, M., and Bell, R. (2013). Sex, cheating, and disgust: enhanced source memory for trait information that violates gender stereotypes. *Memory* 21, 167–181. doi: 10.1080/09658211.2012.713971

Kroneisen, M., Rummel, J., and Erdfelder, E. (2014). Working memory load eliminates the survival processing effect. *Memory* 22, 92–102. doi: 10.1080/09658211.2013.815217

Kroneisen, M., Woehe, L., and Rausch, L. S. (2015). Expectancy effects in source memory: how moving to a bad neighborhood can change your memory. *Psychon. Bull. Rev.* 22, 179–189. doi: 10.3758/s13423-014-0655-9

Küppers, V., and Bayen, U. J. (2014). Inconsistency effects in source memory and compensatory schema-consistent guessing. *Q. J. Exp. Psychol.* 67, 2042–2059. doi: 10.1080/17470218.2014.904914

Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *J. Mem. Lang.* 53, 513–531. doi: 10.1016/j.jml.2005.07.002

Mattarozzi, K., Todorov, A., and Codispoti, M. (2015). Memory for faces: the effect of facial appearance and the context in which the face is encountered. *Psychol. Res.* 79, 308–317. doi: 10.1007/s00426-014-0554-8

Mehl, B., and Buchner, A. (2008). No enhanced memory for faces of cheaters. *Evol. Hum. Behav.* 29, 35–41. doi: 10.1016/j.evolhumbehav.2007.08.001

Minear, M., and Park, D. (2004). A lifespan database of adult facial stimuli. *Behav. Res. Methods Instrum. Comput.* 36, 630–633. doi: 10.3758/BF03206543

Moshagen, M. (2010). multiTree: a computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* 42, 42–54. doi: 10.3758/BRM.42.1.42

Murty, V., FeldmanHall, O., Hunter, L., Phelps, E., and Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *J. Exp. Psychol. Gen.* 145, 548–558. doi: 10.1037/xge0000158

Nash, R. A., Bryer, O. M., and Schlaghecken, F. (2010). Look who's talking! Facial appearance can bias source monitoring. *Memory* 18, 451–457. doi: 10.1080/09658211003742706

Naveh-Benjamin, M., Guez, J., and Marom, M. (2003). The effects of divided attention at encoding on item and associative memory. *Mem. Cogn.* 31, 1021–1035. doi: 10.3758/BF03196123

Nieznański, M. (2013). Effects of resource demanding processing on context memory for context-related versus context-unrelated items. *J. Cogn. Psychol.* 25, 745–758. doi: 10.1080/20445911.2013.819002

Olivola, C. Y., and Todorov, A. (2010). Elected in 100 milliseconds: appearance-based trait inferences and voting. *J. Nonverbal Behav.* 34, 83–110. doi: 10.1007/s10919-009-0082-1

Phillips, P. J., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* 16, 295–306. doi: 10.1016/S0262-8856(97)00070-X

Piovesan, M., and Wengström, E. (2009). Fast or fair? A study of response times. *Econ. Lett.* 105, 193–196.

Rahhal, T., May, C., and Hasher, L. (2002). Truth and character: sources that older adults can remember. *Psychol. Sci.* 13, 101–105. doi: 10.1111/1467-9280.00419

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature* 489, 427–430. doi: 10.1038/nature11467

Rezlescu, C., Duchaine, B., Olivola, C., Chater, N., and Rustichini, A. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE* 7:e34293. doi: 10.1371/journal.pone.0034293

Riefer, D. M., and Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychol. Rev.* 95, 318–339. doi: 10.1007/s00406-010-0151-9

Rule, N., Slepian, M., and Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition* 125, 207–218. doi: 10.1016/j.cognition.2012.06.017

Schmitt, M., Gollwitzer, M., Maes, J., and Arbach, D. (2005). Justice sensitivity: assessment and location in the personality space. *Eur. J. Psychol. Assess.* 21, 202–211. doi: 10.1027/1015-5759.21.3.202

Singmann, H., and Kellen, D. (2013). MPTinR: analysis of multinomial processing tree models in R. *Behav. Res. Methods* 45, 560–575. doi: 10.3758/s13428-012-0259-0

Snodgrass, J. G., and Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *J. Exp. Psychol. Gen.* 117, 34–50. doi: 10.1037/0096-3445.117.1.34

Stahl, C., and Klauer, K. (2007). HMMTree: a computer program for latent-class hierarchical multinomial processing tree models. *Behav. Res. Methods* 39, 267–273. doi: 10.3758/BF03193157

Suzuki, A., and Suga, S. (2010). Enhanced memory for the wolf in sheep's clothing: facial trustworthiness modulates face-trait associative memory. *Cognition* 117, 224–229. doi: 10.1016/j.cognition.2010.08.004

Tinghög, G., Andersson, A., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., et al. (2013). Intuition and cooperation reconsidered. *Nature* 489, 427–430.

Todorov, A. (2008). Evaluating faces on trustworthiness. *Ann. N. Y. Acad. Sci.* 1124, 208–224. doi: 10.1196/annals.1440.012

Todorov, A., Olivola, C., Dotsch, R., and Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* 66, 519–545. doi: 10.1146/annurev-psych-113011-143831

Todorov, A., Pakrashi, M., and Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Soc. Cogn.* 77, 813–833. doi: 10.1521/soco.2009.27.6.813

Todorov, A., and Uleman, J. (2002). Spontaneous trait inferences are bound to actors' faces: evidence from a false recognition paradigm. *J. Pers. Soc. Psychol.* 83, 1051–1065. doi: 10.1037/0022-3514.83.5.1051

Todorov, A., and Uleman, J. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *J. Exp. Soc. Psychol.* 39, 549–562. doi: 10.1016/S0022-1031(03)00059-3

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57. doi: 10.1086/406755

Vachon, F., Labonté, K., and Marsh, J. E. (in press). Attentional capture by deviant sounds: a non-contingent form of auditory distraction? *J. Exp. Psychol. Learn. Mem. Cogn.*

van 't Wout, M., and Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition* 108, 796–803. doi: 10.1016/j.cognition.2008.07.002

Verkoeijen, P. P. J. L., and Bouwmeester, S. (2014). Does Intuition Cause Cooperation? *PLoS ONE* 9:e96654. doi: 10.1371/journal.pone.0096654

Volstorf, J., Rieskamp, J., and Stevens, D. R. (2011). The good, the bad, and the rare: memory for partners in social interactions. *PLoS ONE* 6:e18945. doi: 10.1371/journal.pone.0018945

Willis, J., and Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* 17, 592–598. doi: 10.1111/j.1467-9280.2006.01750.x

Conflict of Interest Statement: 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.

Copyright © 2016 Mieth, Bell and Buchner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 December 2016
doi: 10.3389/fpsyg.2016.02037

Trustworthy Tricksters: Violating a Negative Social Expectation Affects Source Memory and Person Perception When Fear of Exploitation Is High

Philipp Süssenbach[1], Mario Gollwitzer[1], Laura Mieth[2], Axel Buchner[2] and Raoul Bell[2]*

[1] Department of Psychology, Philipps-University Marburg, Marburg, Germany, [2] Department of Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

People who are high in victim-sensitivity—a personality trait characterized by a strong fear of being exploited by others—are more likely to attend to social cues associated with untrustworthiness rather than to cues associated with trustworthiness compared with people who are low in victim-sensitivity. But how do these people react when an initial expectation regarding a target's trustworthiness turns out to be false? Results from two studies show that victim-sensitive compared with victim-insensitive individuals show enhanced source memory and greater change in person perception for negatively labeled targets that violated rather than confirmed negative expectations (the "trustworthy trickster"). These findings are in line with recent theorizing on schema inconsistency and expectancy violation effects in social cognition and with research on the different facets of justice sensitivity in personality psychology.

Keywords: expectancy violation, fear of exploitation, memory, trustworthiness, victim sensitivity

OPEN ACCESS

Edited by:
Bernhard Hommel,
Leiden University, Netherlands

Reviewed by:
Sascha Topolinski,
University of Cologne, Germany
Lindsay R. L. Larson,
Georgia Southern University, USA

****Correspondence:***
Philipp Süssenbach
suessenp@staff.uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *05 May 2016*
Accepted: *15 December 2016*
Published: *27 December 2016*

Citation:
Süssenbach P, Gollwitzer M, Mieth L, Buchner A and Bell R (2016) Trustworthy Tricksters: Violating a Negative Social Expectation Affects Source Memory and Person Perception When Fear of Exploitation Is High. Front. Psychol. 7:2037. doi: 10.3389/fpsyg.2016.02037

INTRODUCTION

Cooperation between individuals requires mutual trust. If person A is in dire straits and asks person B to lend him or her some money, then A should trust that B will not exploit A's state of emergency, and B should trust that A will eventually pay the money back. Neither person can be sure that this is actually the case; this makes the described exchange fundamentally uncertain. This is the paradox of trust (Yamagishi, 2001): the more uncertain a situation, the more trust is required, but—at the same time—the more difficult it is to decide whether one's interaction partner is actually trustworthy.

Humans have a fundamental aversion to being exploited by others. However, this aversion is stronger for some people than for others: People who are victim-sensitive harbor a latent fear of being exploited and react particularly strongly toward experiences of unfairness (Gollwitzer et al., 2005, 2013; Gollwitzer and Rothmund, 2009). Because the possibility of exploitation is aversive and present in many contexts, people are well advised to trust others only when there is reason to do so. Stated differently, whenever a specific social situation entails cues suggesting that one's interaction partner is not trustworthy, then trust becomes riskier and, thus, less likely. It is therefore highly functional (in particular for people aversive to exploitation) to attend to cues that are informative about another person's untrustworthiness, and research shows that people actually do use these cues before they make a trustworthiness decision (e.g., Zebrowitz and Montepare, 2008;

Gollwitzer et al., 2009; Todorov et al., 2009). Such cues include current behavioral cues such as the things a person says, the way he or she looks, or their facial expression, as well as information about a person's past behavior, reputation, or background. Negative social labels, in particular, can be used to quickly form an impression about an interaction partner, and to guide social behaviors. However, these social labels may fail to do justice to each individual. A person who is said to be a trickster (i.e., a negative social label) is possibly regarded as untrustworthy, but may turn out to actually be a very nice and helpful person.

Therefore, it can be considered functional to attend to negative social labels because they provide a quick orientation in a complex social environment, but it is also important to remember information that is inconsistent with these labels, and to integrate it in one's judgment. The present study examines what happens when an untrustworthiness-related cue turns out to be invalid: would victim-sensitive participants with their strong fear of exploitation be able to remember that a trickster turns out to be trustworthy? Or would they show an inflexible memory bias for untrustworthy behavior?

In Study 1 we will show that, perhaps counterintuitively so, victim-sensitive compared with victim-insensitive individuals indeed have a memory advantage for the trustworthy (but not for the untrustworthy) trickster. In Study 2 we will show that victim sensitivity also has an asymmetric effect on people's person perception: Victim-sensitive compared with victim-insensitive participants update their trustworthiness perceptions about the trustworthy trickster more strongly than about the untrustworthy trickster, whereas the updating of another type of expectancy-inconsistent target (e.g., the untrustworthy scientist) is not influenced by victim sensitivity. These findings are incompatible with the notion of a "cheater detection module" (Cosmides, 1989), but they can be well explained by modern schema inconsistency and expectancy violation theories, as will be described in the following.

VICTIM SENSITIVITY AND EXPECTANCY VIOLATION

Victim sensitivity is a self-directed concern for justice characterized by a fear of being exploited. It predicts less pro- and more anti-social behavior (Gollwitzer et al., 2005). Past research has demonstrated that victim sensitivity is a highly stable personality trait (Schmitt et al., 2005) and has documented its location in the personality space (for the relationships with jealousy, just-world beliefs, or Machiavellianism, see Schmitt et al., 2005; for the relationships with the Big Five personality traits, see Schmitt et al., 2010). According to the Sensitivity to Mean Intentions (SeMI) model (Gollwitzer and Rothmund, 2009), victim-sensitive individuals are specifically sensitive to contextual cues that are associated with meanness, recklessness, and untrustworthiness. In social dilemma situations—that are typically characterized by some degree of uncertainty concerning one's partners' intentions—victim-sensitive individuals expect to be exploited and thus tend to defect (Gollwitzer et al., 2009). Hence, some of the uncooperative and anti-social behaviors displayed by people high in victim sensitivity can be understood as a means to protect themselves from (assumed) victimization.

Whereas previous studies have primarily focused (1) on the cognitive schemas (i.e., untrustworthiness expectations) that victim-sensitive individuals apply in social situations and (2) on the behavioral consequences of victim sensitivity in these situations, the present study will be the first to investigate the effect of being confronted with schema-incongruent information on source memory and person perception. In other words, the following questions will be addressed by the present studies: Do expectancy violations have a source memory advantage for victim-sensitive individuals? Are victim-sensitive compared with victim-insensitive individuals more influenced by a violation of positive or negative expectations? And do they update their cognitive schemas accordingly?

CHEATER DETECTION AND EXPECTANCY VIOLATION

According to evolutionary psychologists, the analysis of evolutionary pressures is essential for understanding how the human mind works. One such pressure is the maintenance of social exchange within larger groups of non-kin, as cooperation between unrelated individuals is prone to be exploited by cheaters. Thus, the ability to identify and remember people who cannot be trusted is considered particularly adaptive (Cosmides and Tooby, 1989, 2005); hence, many authors argue in favor of the existence of a specialized cognitive module devoted to the detection of, and memory for, cheaters. Indeed, there is some empirical evidence that supports the assumption of a specialized "cheater detection module" (Mealey et al., 1996; Oda, 1997).

Other findings, however, speak against the existence of a specialized cheater detection module and are in favor of more general mechanisms (Bell and Buchner, 2012): in the area of memory research, participants usually show enhanced memory for the violation of both positive *and* negative expectancies (Barclay, 2008; Bell et al., 2010, 2012, 2015; Suzuki and Suga, 2010; Mieth et al., 2016). Thus, human memory is indeed adaptive, but even more strongly than suggested. In line with an evolutionary account, memory for cheaters is quite good in contexts in which cooperation is the norm and cheating is unexpected. However, this pattern flips when the context changes participants' expectations. For example, in cooperation games with very low cooperation rates trustworthy individuals are particularly well remembered (Barclay, 2008; Bell et al., 2010). These findings suggest that enhanced cheater memory is best explained in terms of an expectancy violation or schema inconsistency account.

The schema-plus-tag model (Graesser and Nakamura, 1982), for example, states that memory discrimination for schema-consistent information is poor because schema-consistent information is always produced at test, whether it was actually present at encoding or not. According to this model, memory discrimination for the untrustworthy behavior of a trickster would be poor because untrustworthiness is already part of the negative stereotype of a trickster, and is copied into the memory trace (guessed), regardless of whether it was actually present

at encoding or not. Schema-*inconsistent* information (e.g., a trustworthy behavior of a trickster), in contrast, is stored in memory in the form of tags.

It is therefore an open question how memory accuracy for different social targets is affected by victim sensitivity. How do victim-sensitive individuals process information that violates or confirms their negative social expectations about a particular target? Based on what is known so far, two patterns of results are conceivable. On the one hand, if the perception of, and memory for, cheaters was driven by experienced negativity (as the "cheater detection" literature would suggest), then individuals who are victim-sensitive should be more influenced in their judgment and have enhanced memory for cheaters (as they experience stronger emotional reactions toward learning that someone is truly untrustworthy) than individuals who are less victim-sensitive.

On the other hand, if the perception of, and memory for, cheaters was driven rather by schema inconsistency than by negativity (as the current state of memory research suggests), then a different pattern should be expected. Untrustworthy behavior is already a part of the negative stereotypes associated with negative social labels. Therefore it does not change one's attitude toward that target and does not have to be remembered separately. Learning that a dubious target has acted truly trustworthily, however, comes as a much bigger surprise to people who fear exploitation than to those who do not, as the former hold much more negative expectations toward such targets in the first place. Thus, if individuals' memory is particularly good for schema-inconsistent behaviors, then a stronger fear of exploitation should cause better source memory for targets *violating* negative expectancies than for targets *confirming* negative expectancies. Likewise, we would expect victim sensitivity to predict a greater change in the perception of targets violating negative expectancies than in the perception of targets confirming negative expectancies. Put more bluntly, victim-sensitive compared with victim-insensitive individuals should remember the untrustworthy behavior of a trickster particularly poorly because the untrustworthiness of this target is already part of the negative stereotype while the atypical trustworthy information about the trickster should be remembered particularly well. The current state of research on memory for cheaters and non-cheaters suggests that enhanced memory is most likely to be driven by schema inconsistency (see Bell and Buchner, 2012); thus, we consider the latter pattern of results to be more likely.

RESEARCH OVERVIEW

The present research aims to investigate how victim-sensitive individuals compared with victim-insensitive individuals perceive and memorize targets with negative or positive social labels when these targets supposedly did something that was inconsistent with their respective label. Based on the theorizing presented above, it is predicted that victim-sensitive individuals hold more negative expectations toward targets associated with negative social labels, which should influence the classification of these targets in a subsequent memory test in two ways. First, victim-sensitive compared with victim-insensitive individuals should rely heavily on their biased expectations when memory is not available. In consequence, they should show a more pronounced bias toward guessing that targets with negative social labels were previously associated with untrustworthy behavior. Second, regarding memory accuracy, it is predicted that victim sensitivity is associated with enhanced memory for *violations,* but not for *confirmations,* of negative expectations. Third, it is predicted that victim-sensitive individuals are more likely to update their trustworthiness perceptions for negative expectancy violations, but not for negative expectancy confirmations.

Predictions regarding violations of *positive* expectations are less straightforward. Past research has shown that both victim-sensitive and victim-insensitive individuals react similarly to cues of trustworthiness (Gollwitzer et al., 2012). Hence, their expectations regarding targets with positive social labels (i.e., "scientist") are not expected to differ. However, violations of these positive expectations might pose a greater threat to victim-sensitive individuals, thereby affecting memory and person perception more strongly. Thus, whereas initial expectations toward these targets should not be influenced by victim sensitivity (and thus a pure expectancy violation account would not predict an effect of victim sensitivity on memory and perception for such cases), it is conceivable that the trustworthiness violation itself is stronger for people high in victim sensitivity (which would imply an effect of victim sensitivity also in cases of a violation of positive expectations).

These hypotheses were tested in two studies. Study 1 examined the influence of victim sensitivity on source memory. To that end, participants viewed faces that were accompanied by a positive (e.g., scientist, firefighter) or negative social label (e.g., trickster, former prisoner). After a short delay this information was complemented with a behavioral description that represented either prosocial (i.e., trustworthy) or antisocial (i.e., untrustworthy) behavior. After viewing these profiles, participants completed a surprise source memory test in which they viewed faces and indicated whether a face had been presented before, and, if so, whether it was paired with trustworthy or untrustworthy behavior. It is predicted that participants high in victim sensitivity compared to participants low in victim sensitivity have more negative social expectations toward targets with negative social labels, which is reflected in a bias toward guessing (i.e., in the absence of memory about the correct behavior) that faces with negative labels are associated with untrustworthy behavior. This finding would be consistent with prior research showing that victim-sensitive individuals rely particularly strongly on untrustworthiness cues (Gollwitzer et al., 2012). Importantly, these negative social expectations should result in particularly good source memory for the *violation* of *negative* expectations; that is, for negatively labeled targets who displayed trustworthy (compared to untrustworthy) behavior.

In Study 2, participants' perceptions of the targets' trustworthiness were examined before and after they learned about the trustworthy or untrustworthy behavior of the positively or negatively labeled targets. Importantly, an experimental

manipulation was included that aimed at making fear of exploitation salient to demonstrate that differential effects of victim sensitivity are indeed causally attributable to differences in people's victim sensitivity. Thus, it is predicted that victim-sensitive participants in the fear of exploitation condition harbor more negative initial expectations toward targets with negative social labels than victim-sensitive participants in the control condition or victim-insensitive participants. Given that higher victim sensitivity should be associated with greater schema violation regarding dubious targets who show trustworthy behavior (compared to dubious targets who show untrustworthy behavior), changes in perceived trustworthiness should be stronger for "trustworthy tricksters" than for "untrustworthy tricksters" (i.e., selective updating). Importantly, these effects should be more pronounced in the fear of exploitation condition than in the control condition.

STUDY 1

Method

Participants

Participants were recruited from an undergraduate student pool of a large German university. One data set was removed because it turned out later the participant had participated twice in the experiment. The remaining sample consisted of 104 students (68 women; $M_{Age} = 24$, $SD_{Age} = 4$).

Materials

Ninety pictures (512 × 768 pixels) of frontal male faces with a neutral facial expression were selected from the FERET database (Phillips et al., 1998). We only used faces that had received neutral ratings of facial trustworthiness in a norming study ($M = 3.28$ on a scale ranging from 1 [not at all trustworthy] to 6 [very trustworthy]; $SD = 0.22$).

In a separate norming study, 15 participants ($M_{Age} = 24$, $SD_{Age} = 2$) rated the trustworthiness of 194 social labels using a scale ranging from 1 (not at all trustworthy) to 6 (very trustworthy). Out of these, 45 positive labels with a mean trustworthiness of 4.43 ($SD = 0.25$) and 45 negative labels with a mean trustworthiness of 1.92 ($SD = 0.45$) were selected as stimulus material. Examples for positive labels are "scientist," "professor," "firefighter," and "ambulance driver;" examples for negative labels were "trickster," "Satanist," "former prisoner," or "gang member."

In yet another norming study, ($N = 40$, $M_{Age} = 28$, $SD_{Age} = 10$), behavioral descriptions of trustworthy and untrustworthy behaviors were rated on a scale ranging from −3 (very untrustworthy) to +3 (very trustworthy). The 25 descriptions of untrustworthy behavior had a mean rating of −2.22 ($SD = 0.42$), and the 25 descriptions of trustworthy behavior had a mean rating of +1.89 ($SD = 0.45$). An example for untrustworthy behavior is "He exploits the trust of older people and steals valuable items from their apartments." An example of trustworthy behavior is "On his way home he once risked his life to rescue a kid that fell into a frozen pond."

Procedure

In the encoding phase, participants saw 50 male faces. The faces were randomly assigned to 25 negative and 25 positive labels. Each trial started with the presentation of a face. Below the facial photograph, a label was presented (e.g., "F. D. is a scientist," or "S. D. is a trickster"). After 4.5 s, face and label were complemented by a behavioral description. The behavioral descriptions were randomly selected with the restriction that the negative labels were paired with 15 untrustworthy and 10 trustworthy descriptions and the positive labels were paired with 15 trustworthy and 10 untrustworthy descriptions. Participants were required to rate the likeability of the person, and then initiated the next trial by clicking on a "continue" button. Negative and positive social labels were paired with more valence-congruent in comparison to valence-incongruent behaviors because people's negative or positive stereotypes should not be blatantly disconfirmed by a high proportion of schema-inconsistent pairings in the encoding phase.

Immediately after the encoding phase, participants received instructions for a surprise source memory test, in which 80 facial photographs were presented in a random order. Each face was accompanied by a social label (i.e., "scientist," "trickster"). Half of the faces had been presented in the encoding phase. Ten faces with negative labels had been described as untrustworthy, 10 faces with negative labels had been described as trustworthy, 10 faces with positive labels had been described as untrustworthy, and 10 faces with positive labels had been described as trustworthy. Of the 40 new faces, 20 were accompanied by negative labels, and 20 were accompanied by positive labels. The faces and labels were randomly selected to be presented in either the encoding phase or test phase. Faces and labels were randomly assigned to conditions.

Each test trial started with the presentation of a face with a label. After a 1.5 s interval, the likeability rating scale appeared (ranging from 1 [not at all likeable] to 6 [very likeable]). After rating the person's likeability, the participants were asked to indicate whether the face was old or new (had been presented during the encoding phase or not). When the person had been classified as old, participants were asked to indicate whether the person was accompanied by a trustworthy or an untrustworthy behavior description during the encoding phase. After the test phase, participants completed a paper-and-pencil version of Schmitt et al.'s justice sensitivity questionnaire (Schmitt et al., 2010). Victim sensitivity was assessed with 10 items (Cronbach's $\alpha = 0.82$). Example items are "It bothers me when others receive something that ought to be mine" or "It makes me angry when others receive a reward that I have earned." Each item was rated on a 6-point response scale ranging from 1 (not at all true) to 6 (absolutely true).

Design

A multinomial model was used to distinguish between old-new recognition, source memory, and source guessing. Given an estimated small effect of $\varpi = 0.04$ (estimated on the basis of pilot studies), $\alpha = 0.05$, and 80 answers in the source memory test, an N of 104 is sufficient to detect an effect with a power of $1 - \beta = 0.95$. Power was calculated using G*Power (Faul et al., 2007).

Results and Discussion

Source Memory and Source Guessing

A multinomial source monitoring model (Bayen et al., 1996; see **Figure 1**) was used to distinguish between guessing, source memory, and old–new recognition. This model is well validated (Bayen et al., 1996; Erdfelder et al., 2009), and has been used in previous studies to disentangle the effects of schema (in-) consistency on source guessing and source memory (Bayen and Kuhlmann, 2011; Bell et al., 2015). The first tree represents the processing tree of a face-label pair that was paired with a description of untrustworthy behavior during the encoding phase. With probability D_C, participants recognize the face-label combination as old. With the conditional probability d_C, source memory, that is, memory for the association of the face-label combination with an untrustworthy behavior description, is correctly remembered, in which case the participant is able to correctly classify the person as a cheater. With the complementary probability 1 – d_C, the participant has no source memory. In this case, the participant has to guess, with probability g, that the person was described as a cheater, or, with probability 1 – g, that the person was associated with trustworthy behavior. With probability 1 – D_C, the participant fails to recognize the face-label combination as old, in which case participants guess, with probability b, that the item is old, or, with probability 1 – b, that the item is new. If the item is guessed to be old, participants further guess, with probability g, that the person was described as a cheater, or, with probability 1 – g, that the person was described as trustworthy.

The other two trees represent the processing of face-label combinations that were associated with trustworthy behavior descriptions and face-label combinations that were new (only presented at test), respectively. Model 5d of Bayen et al.'s (1996) taxonomy of identifiable submodels, which includes the restriction $D_C = D_T = D_{New}$, was used[1]. Two sets of the processing trees displayed in **Figure 1** are needed for the analysis of the present data set, one for faces with negative labels, and one for faces with positive labels. Parameter estimations and goodness-of-fit tests were performed using multiTree (Moshagen, 2010).

The base model provided a good fit to the data, $G^2(2) = 0.05$, $p = 0.97$. When no source memory was available, participants had a higher probability of guessing that a person had been described as a cheater when the social label was negative than when the label was positive, $\Delta G^2(1) = 6.72$, $p < 0.01$. The estimates of the source guessing parameter g and of the source memory parameter d are reported in **Tables 1** and **2,** respectively.

Consistent with previous studies (Mieth et al., 2016), participants showed an asymmetric expectancy violation effect. For faces with negative labels, source memory was better for the trustworthy behaviors in comparison to untrustworthy behaviors, $\Delta G^2(1) = 4.74$, $p = 0.03$. For faces with positive labels, source memory did not differ as a function of behavior type, $\Delta G^2(1) = 0.70$, $p = 0.40$. This is clearly at odds with the assumption of a negativity advantage, and suggests that participants flexibly shifted their attention to information that was unexpected and, therefore, most informative for them.

Victim Sensitivity

Next, we wanted to know how guessing and source memory were affected by victim sensitivity. As reasoned in the introduction, *a priori* it seemed possible that victim-sensitive persons would show a particularly inflexible memory advantage for cheating (Gollwitzer et al., 2012). The schema violation explanation, however, would suggest that victim-sensitive persons would show stronger negative social expectations based on the negative social labels and, thus, stronger schema violation effects for trustworthy tricksters (whereas for untrustworthy scientists no difference in initial expectations due to victim sensitivity was expected).

To analyze the influence of victim sensitivity, we followed the exact same procedure as Mieth et al. (2016). Victim sensitivity was dichotomized at its sample median (i.e., 2.9). Fifty-four participants constituted the low victim sensitivity group, 50 constituted the high victim sensitivity group. These data were analyzed together using separate model trees for individuals with high and low victim sensitivity. The base model (incorporating the same restrictions as the base model reported above) fit the data well, $G^2(4) = 1.38$, $p = 0.85$. More importantly, and in line with our expectations, participants in the high victim sensitivity group had a bias toward guessing that targets with negative labels were more likely to be cheaters than targets with positive labels, $\Delta G^2(1) = 6.13$, $p = 0.01$ (see **Table 3**).

When compared against a neutral baseline of guessing with 0.50 that the target was either described as a cheater or as a trustworthy person, participants in the high victim sensitivity group had a bias toward guessing that a target with a negative label was associated with untrustworthy behavior, $\Delta G^2(1) = 5.01$, $p = 0.03$, but no bias toward guessing that a target with a positive label was associated with trustworthiness, $\Delta G^2(1) = 1.38$, $p = 0.24$. Thus, the guessing bias of victim-sensitive participants was stronger in the negative than in the positive direction. Participants in the low victim sensitivity group, in contrast, showed no such bias. Their tendency toward guessing that a target was described as a cheater (when no source memory was available) was not significantly affected by the negative or positive social label, $\Delta G^2(1) = 1.63$, $p = 0.20$.

In addition, participants in the high victim sensitivity group had enhanced source memory for violations of their label-based expectations—that is, for descriptions of trustworthiness in comparison to descriptions of untrustworthiness when the targets were associated with negative labels, $\Delta G^2(1) = 5.54$, $p = 0.02$ (see **Table 4**). As in the global analysis, there was no difference between source memory for untrustworthy and trustworthy descriptions when positive labels were used, $\Delta G^2(1) = 0.35$, $p = 0.55$. Participants in the low victim sensitivity group showed no such schema inconsistency advantage in source memory. In fact, there was no difference between untrustworthy and trustworthy descriptions, regardless of

[1]The hypothesis that $D_C = D_T$ is supported by the available literature (e.g., Barclay and Lalumière, 2006; Buchner et al., 2009), as well as by the present results. Furthermore, the statistical comparisons do not change when using a base model that does not make this assumption and simply assumes that $D_{New} = (D_C + D_T)/2$ (as in Bell et al., 2015).

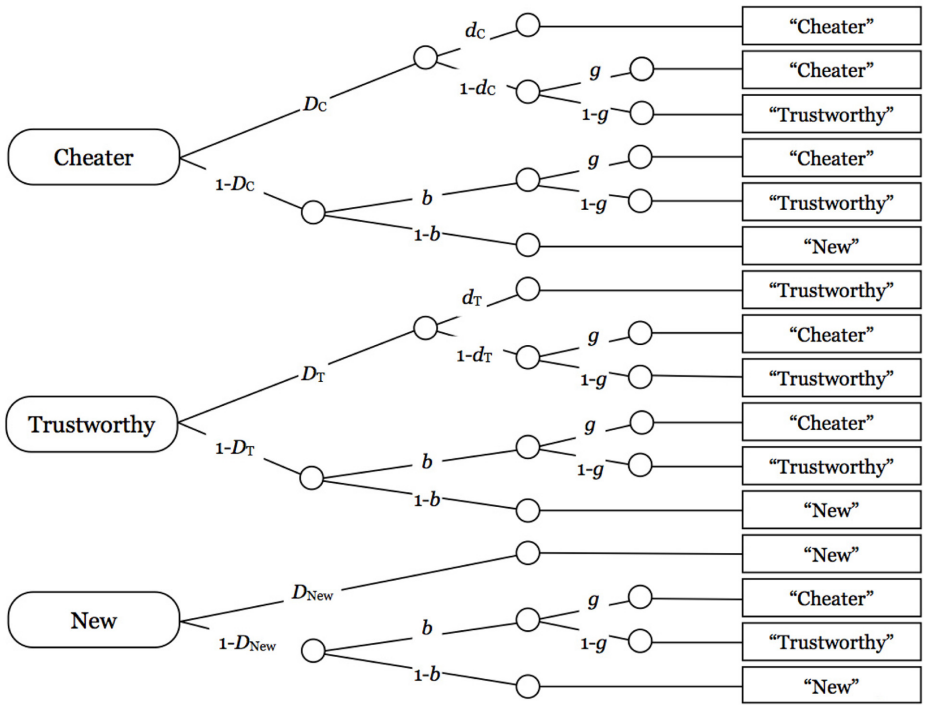


FIGURE 1 | Bayen et al.'s (1996) source memory model, adapted for the present study. Rounded rectangles on the left side represent the stimulus persons (cheaters, trustworthy persons, and new persons). Letters along the branches represent the model parameters. *D*: probability that a person (a face with a label) is correctly recognized as old or new. *d*: conditional probability that the context (untrustworthy or trustworthy behavior) is correctly remembered. *b*: conditional probability of guessing that a person has been presented during the encoding phase. *g*: conditional probability of guessing that a person has previously been associated with untrustworthy behavior.

TABLE 1 | Parameter estimates of the guessing bias parameter *g* representing the conditional probability of guessing that the person was a cheater rather than a trustworthy person as a function of label (Study 1).

Label	Parameter estimate	(*SE*)	[0.95 Confidence Interval]
Negative	0.61	(0.05)	[0.51 –0.71]
Positive	0.43	(0.04)	[0.35 –0.52]

TABLE 2 | Parameter estimates of the source memory parameter *d* as a function of label and behavior (Study 1).

Label	Behavior	Parameter Estimate	(*SE*)	[0.95 Confidence Interval]
Negative	Untrustworthy	0.27	(0.11)	[0.05 –0.49]
Negative	Trustworthy	0.57	(0.05)	[0.48 –0.67]
Positive	Untrustworthy	0.38	(0.06)	[0.25 –0.50]
Positive	Trustworthy	0.48	(0.07)	[0.34 –0.62]

TABLE 3 | Parameter estimates of the guessing bias parameter *g* representing the conditional probability of guessing that the person was a cheater rather than a trustworthy person as a function of label and victim sensitivity (Study 1).

Label	Parameter Estimate	(*SE*)	[0.95 Confidence Interval]
High victim sensitivity			
Negative	0.67	(0.07)	[0.53 –0.82]
Positive	0.42	(0.06)	[0.30 –0.55]
Low victim sensitivity			
Negative	0.56	(0.07)	[0.42 –0.70]
Positive	0.44	(0.06)	[0.32 –0.56]

whether the labels were negative, $\Delta G^2(1) = 0.67$, $p = 0.41$, or positive, $\Delta G^2(1) = 0.33$, $p = 0.56$.

In summary, Study 1 suggests that victim-sensitive compared with victim-insensitive persons have stronger negative expectations toward people associated with negative social labels, as reflected in a bias toward guessing that targets with negative labels have been associated with negative social behaviors. This finding nicely fits with prior research demonstrating that victim-sensitive individuals are more likely to use untrustworthiness cues than victim-insensitive individuals (Gollwitzer et al., 2012).

As hypothesized, the stronger initial influence of untrustworthiness cues led to an "ironic" schema inconsistency effect in the source memory of victim-sensitive individuals: behavior that was inconsistent with the negative labels was particularly well remembered as evidenced in better memory for negative targets that violated rather than confirmed negative expectations. Thus, victim sensitivity seems to be associated with a reliance on negative expectations. This reliance on negative expectations resulted in a schema-consistent guessing bias for

TABLE 4 | Parameter estimates of the source memory parameter *d* as a function of label, behavior and victim sensitivity (Study 1).

Label	Behavior	Parameter estimate	(*SE*)	[0.95 Confidence Interval]
High victim sensitivity				
Negative	Untrustworthy	0.08	(0.23)	[0.00 −0.53]
Negative	Trustworthy	0.61	(0.06)	[0.50 −0.73]
Positive	Untrustworthy	0.37	(0.09)	[0.19 −0.54]
Positive	Trustworthy	0.48	(0.11)	[0.27 −0.68]
Low victim sensitivity				
Negative	Untrustworthy	0.39	(0.12)	[0.16 −0.62]
Negative	Trustworthy	0.54	(0.08)	[0.39 −0.68]
Positive	Untrustworthy	0.39	(0.09)	[0.22 −0.56]
Positive	Trustworthy	0.49	(0.09)	[0.30 −0.67]

negative labels. Moreover, the reliance on negative expectations of victim-insensitive individuals enhanced source memory for information that was inconsistent with these negative labels relative to information that was consistent therewith (indeed source memory for expectancy confirming negative targets was extremely low). Thus, whereas victim-sensitive individuals would probably prefer to remember targets well who behaved negatively (i.e., "the unstrustworthy trickster" or "the unstrustworthy scientist"), they are likely – due to strong initial negative expectations regarding targets with a negative social label – to remember those targets who surprised them by displaying positive behavior.

One limitation of the present data analytic procedure needs to be mentioned. Whereas the multinomial model is necessary to distinguish between memory and guessing, it does not allow for a direct test of interaction effects. Thus, testing "interaction effects" requires running separate models in different subgroups. The same approach has been used in previous source memory studies (Bayen and Kuhlmann, 2011; Bell et al., 2015; Mieth et al., 2016). Our present results suggest that being high in victim sensitivity is not associated with better source memory for untrustworthy behavior in general; rather, the results are more in line with the idea that victim-sensitive individuals rely on their negative schemata when guessing, and remember information that violates their negative expectations. This evidence, however, is indirect because (1) the analysis capitalized on group differences in victim sensitivity, but did not experimentally manipulate fear of exploitation, and (2) the dependent variable in Study 1 focused on the outcome of an expectancy violation, but did not measure intraindividual changes in the perception of trustworthiness after an expectancy-inconsistent vs. expectancy-consistent behavioral description has been provided.

To address these limitations, it seemed necessary to investigate the influence of victim sensitivity on the effects of violations of positive and negative expectations more directly. In Study 2, therefore, participants' person perception rather than their source memory was examined in response to expectancy-congruent vs. expectancy-incongruent information about the targets. To foster the argument that differences in person perception are indeed causally attributable to victim sensitivity as a personality trait reflective of a latent fear of being exploited, an experimental manipulation was introduced to activate victim sensitivity.

STUDY 2

Method

Participants and Design

Assuming $\alpha = \beta = 0.05$ and an effect of $\phi^2 = 0.25$ regarding the condition × victim sensitivity interaction on trustworthiness perception in a multiple regression analysis, 54 observations were needed. With a final sample size of 60, an effect of $\phi^2 = 0.22$ could be detected. The final sample consisted of 60 students (51 women, 8 men, 1 non-response) of a German university ($M_{Age} = 22$, $SD_{Age} = 5$). Participants were randomly assigned to one of two conditions, the exploitation condition ($n = 29$) or the control condition ($n = 31$; see below for details).

Materials

Fear of exploitation manipulation

Participants first read a short scenario. Their role in the scenario varied depending on experimental conditions. Participants in the exploitation condition were asked to imagine that they would have to give a presentation with two fellow students in a university course, and that they would receive a grade for their presentation which was very important to them. What follows is a summary of the scenario in the exploitation condition:

> *Your presentation is coming up soon and you and the two other students agree to meet in the department library after the end of the course to start with the literature search. However, the other two do not show up, and you are forced to look for the literature on your own. You eventually end up preparing the presentation all by yourself, although you tried to contact the two other students. The day before the presentation, the two others suddenly contact you and ask what their part would be in the presentation. They excuse themselves and say that they had been busy. The presentation itself works out well until the lecturer assigns grades. Your two fellow students receive better grades than you do although they barely invested anything in the presentation, and, to make matters worse, they do not object to this unjust grading. You are stupefied by the behavior of your fellow students and you feel exploited and treated unfairly.*

Participants in the control condition read the same scenario, but from the perspective of an observer. Both scenarios were equal in length and varied only with regard to whether participants imagined that the event befell them (exploitation condition) or someone else (control condition). Thus, the exploitation condition should activate participants' victim sensitivity (via imagining being exploited), whereas the control condition should not have this effect given that someone else (but not oneself) is the victim of injustice. The latter might activate participants' observer sensitivity (Schmitt et al., 2005), but not their victim sensitivity.

Comprehension and manipulation check

Participants completed three items that assessed whether they had difficulties understanding the scenario ("I read the text with

full concentration," $M = 5.70$, $SD = 0.53$; "I found it easy to read the text," $M = 5.88$, $SD = 0.37$; "I can describe the content of the text," $M = 5.73$, $SD = 0.48$; from 1 [not at all true] to 6 [absolutely true]). As a manipulation check, participants responded to three items that assessed moral outrage and anger in response to the situation described in the scenario ("The situation described in the text makes me upset;" "The situation described in the text makes me angry;" "The situation described in the text bothers me;" from 1 [not at all true] to 6 [absolutely true]; $M = 4.92$, $SD = 0.88$, $\alpha = 0.88$).

Trustworthiness perceptions

Next, participants viewed 24 male faces (faces, labels, and behavioral descriptions were taken from the same sources as in Study 1). Half of those were accompanied by positive social labels; the others were accompanied by negative social labels. Trustworthiness was assessed with two items ("How trustworthy is this person?" and "How likeable is this person?" from 1 [not at all] to 6 [very] that was always presented with a filler item ("How competent is this person?" from 1 [not at all] to 6 [very]).

After rating the targets' trustworthiness for the first time (T1), participants viewed the same targets a second time (T2; in a different order). This time, the targets came with a behavioral description (trustworthy vs. untrustworthy behavior); these descriptions were also taken from the same sources as in Study 1. Nine faces with negative labels were paired with untrustworthy behavior, nine faces with positive labels were paired with trustworthy behavior, three faces with negative labels were paired with trustworthy behavior, and three faces with positive labels were paired with untrustworthy behavior.

Victim sensitivity

Finally, participants' victim sensitivity was assessed with the same 10-item scale as in Study 1 (Schmitt et al., 2010; $M = 4.23$, $SD = 0.69$, $\alpha = 0.80$). Participants' victim sensitivity did not differ as a function of the experimental manipulation, $t(58) = 0.75$, $p = 0.45$.

Results and Discussion

Manipulation Check

To test whether the experimental manipulation was successful, a hierarchical regression analysis was conducted with condition (0 = control, 1 = exploitation), victim sensitivity (mean-centered), and the condition × victim sensitivity interaction (entered in a second step) as predictor variables, and moral outrage about the scenario as dependent variable. Neither victim sensitivity, $B = 0.28$, $t(57) = 1.72$, $p = 0.09$, nor condition, $B = 0.19$, $t(57) = 0.91$, $p = 0.40$, had main effects on moral outrage, but the interaction effect was significant, $B = 0.69$, $t(56) = 2.20$, $p = 0.03$, $\Delta R^2 = 0.074$. We probed this interaction using Hayes's (2013) PROCESS macro. For people high in victim sensitivity (1 *SD* above the sample mean), moral outrage scores were significantly higher in the exploitation than in the control condition, $B = 0.67$, $t(56) = 2.18$, $p = 0.03$, whereas for people low in victim sensitivity (1 *SD* below the sample mean), the difference between the two experimental conditions was not significant, $B = -0.29$, $t(56) = -0.94$, $p = 0.35$. Stated differently, victim sensitivity predicted moral outrage in the exploitation condition, $B = 0.59$, $t(56) = 2.79$, $p = 0.007$, but not in the control condition, $B = -0.10$, $t(56) = -0.44$, $p = 0.66$. Thus, the experimental manipulation was successful in activating victim sensitivity.

Trustworthiness Perceptions at T1

Participants' trustworthiness perceptions were aggregated (1) across the nine expectancy-confirming positive targets (positive label plus trustworthy behavioral description), (2) across the 9 expectancy-confirming negative targets (negative label plus untrustworthy behavioral description), (3) across the three expectancy-violating positive targets (positive label plus untrustworthy behavior), and (4) across the three expectancy-violating negative targets (negative label plus trustworthy behavior). Mean trustworthiness perceptions are displayed in **Table 5**.

The effects of experimental condition and victim sensitivity on trustworthiness perceptions at T1 (i.e., without behavioral descriptions) were tested via hierarchical multiple regression analysis with condition (0 = control, 1 = exploitation), victim sensitivity (mean-centered), and their interaction term (entered in a second step) as predictors. These analyses were conducted for targets with positive and targets with negative labels, respectively. For targets with positive labels, neither the experimental manipulation nor victim sensitivity (or their interaction) influenced trustworthiness perceptions, all $ps > 0.42$. For targets with negative social labels, however, trustworthiness perceptions were significantly affected by victim sensitivity, $B = -0.28$, $t(57) = -2.16$, $p = 0.03$, $\Delta R^2 = 0.071$, condition, $B = -0.33$, $t(57) = -1.90$, $p = 0.06$, $\Delta R^2 = 0.055$, and their interaction, $B = -0.50$, $t(56) = 1.99$, $p = 0.05$, $\Delta R^2 = 0.057$ (see **Figure 2**, left panel). Probing this interaction further showed that people high in victim sensitivity (1 *SD* above the sample mean) gave significantly lower trustworthiness perceptions in the exploitation than in the control condition, $B = -0.68$, $t(56) = -2.78$, $p = 0.007$, whereas for people low in victim sensitivity (1 *SD* below the sample mean), this effect was not significant, $B = 0.01$, $t(56) = 0.03$, $p = 0.97$. Stated differently, victim sensitivity predicted lower trustworthiness perceptions in

TABLE 5 | Descriptive findings for the trustworthiness perceptions at T1 and change in perceived trustworthiness at T2 (Study 2).

Measure		*N* Items	Mean	*SE*
Label	**Behavior**			
Trustworthiness perceptions at T1				
Negative		18	2.62	0.71
Negative		6	2.92	0.86
Positive		18	4.25	0.57
Positive		6	4.16	0.71
Difference score (T2-T1)				
Negative	Untrustworthy	18	−1.37	0.61
Negative	Trustworthy	6	1.59	0.70
Positive	Untrustworthy	6	−2.82	0.78
Positive	Trustworthy	18	0.95	0.41

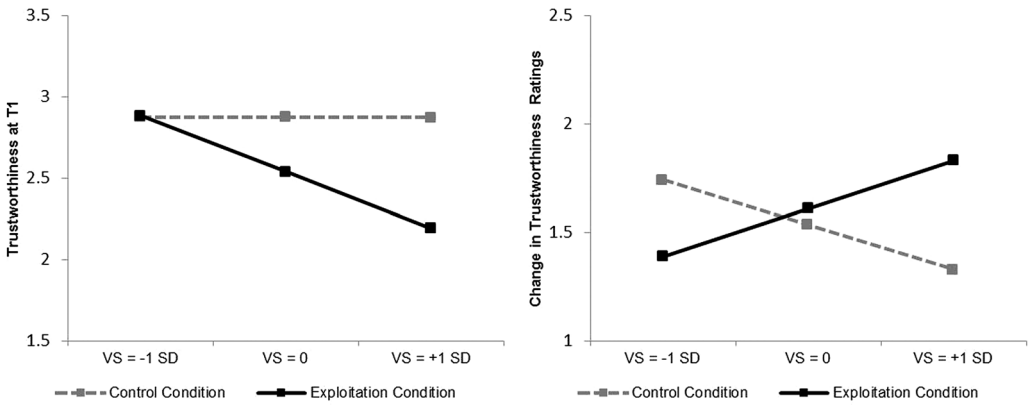


FIGURE 2 | Effect of victim sensitivity and exploitation condition on initial trustworthiness perception (left) and change in trustworthiness (right) for targets with a negative social label and a positive behavioral description. Higher values on change in trustworthiness reflect stronger change toward more trustworthiness. Note that victim sensitivity was included as a continuous predictor in all analyses. Plotting the dependent variables at three levels of victim sensitivity only serves to visualize the findings. VS = victim sensitivity.

the exploitation condition, $B = -0.50$, $t(56) = -2.98$, $p = 0.004$, but not in the control condition, $B = -0.00$, $t(56) = -0.01$, $p = 0.99$.

Changes in Trustworthiness Perceptions

To quantify participants' responsiveness to behavioral descriptions, a difference score was computed by subtracting trustworthiness perceptions at T1 (without behavioral descriptions) from their trustworthiness perceptions at T2 (with behavioral descriptions). Thus, positive values on the difference score reflect an increase in perceived trustworthiness at T2 relative to T1; negative values reflect a decrease. Multiple regression analyses (see above) were conducted to analyze the effect of our predictor variables on expectancy-violating targets.

Regarding expectancy-violating targets with a *positive* social label (e.g., untrustworthy scientists), changes in trustworthiness perceptions were neither predicted by condition, nor by victim sensitivity, nor their interaction, all *p*s > 0.44. However, regarding expectancy-violating targets with a *negative* social label (e.g., trustworthy tricksters), changes in perceived trustworthiness were significantly predicted by the condition × victim sensitivity interaction, $B = 0.62$, $t(56) = 2.37$, $p = 0.02$, $\Delta R^2 = 0.091$ (see **Figure 2**, right panel). Probing this interaction further revealed that people high in victim sensitivity (1 *SD* above the sample mean) showed significantly greater change in perceived trustworthiness in the exploitation compared to the control condition, $B = 0.50$, $t(56) = 1.98$, $p = 0.05$, whereas for people low in victim sensitivity (1 *SD* below the sample mean), this effect was not significant, $B = -0.35$, $t(56) = -1.39$, $p = 0.17$.

In a final step, we tested whether victim-sensitive compared with victim-insensitive individuals indeed update their expectations particularly when a target with a negative social labels turns out to be trustworthy. To do so, a mixed model was performed on participants' absolute change scores regarding their trustworthiness perception of negative targets who violated versus confirmed negative expectations. No restriction was imposed on the covariance matrix and parameters were estimated using full maximum likelihood. This analysis yielded a significant three-way interaction between type of target (confirming vs. violating) × condition × victim sensitivity ($p = 0.008$): Victim-sensitive individuals in the exploitation condition updated their perceptions of negative targets who violated their expectations (see the just reported results of the multiple regression analyses for the direction of this updating), but not of negative targets who confirmed them. Indeed, for negative targets who confirmed negative expectations victim-sensitivity was related to reduced updating of perceptions in the exploitation condition, $B = -0.38$, $t(52) = -2.66$, $p = 0.01$, whereas in the control condition victim sensitivity was unrelated to changes in trustworthiness toward such targets, $B = -0.02$, $t(52) = -0.10$, $p = 0.92$.

Thus, although participants high in victim sensitivity tend to distrust targets with negative social labels initially, they are more likely to selectively update their trustworthiness perception after receiving expectancy-violating information relative to participants low in victim sensitivity. This greater change in perceived trustworthiness seems to reflect that, when fear of exploitation is high, people are particularly responsive to the violation of negative expectations. Moreover, the fact that victim sensitivity predicted greater change in trustworthiness in the condition that activated participants' victim sensitivity is in line with the argument that the observed sensitivity to a violation of negative expectations is indeed causally attributable to differences in participants' fear of exploitation.

GENERAL DISCUSSION

When people are victim-sensitive, they are more receptive toward cues associated with untrustworthiness, such as the interaction partner's facial expression or his or her background. So, when fear of exploitation is high, negative social labels such as "X is a trickster" have a stronger influence on one's trustworthiness

perception of X than a positive social labels such as "X is a scientist." This has been suggested by recent research on victim sensitivity and suspicious cognition (see Gollwitzer and Rothmund, 2009; Gollwitzer et al., 2013, for theoretical reviews). The present study corroborates and extends these findings by asking: what happens if an initial expectation regarding a particular interaction partner is violated, that is, if a "trickster" turns out to be trustworthy rather than untrustworthy? Here, a "cheater detection" account (Cosmides and Tooby, 1989) would predict that people are more likely to attend to (and remember) the latter information. But recent research suggests that memory advantages for cheaters are not as robust as evolutionary psychology thought they would be. Source memory effects can be better conceptualized as expectancy-violation effects than as cheater detection effects (Bell and Buchner, 2012). Thus, it was hypothesized that participants with a fear of exploitation would not show enhanced memory accuracy for the untrustworthy behavior of the trickster because untrustworthiness is already part of their negative stereotype of a trickster. In contrast, they should be more influenced by information that contradicts their initial (negative) expectation. This asymmetric effect should manifest in (a) better source memory and (b) increased changes in trustworthiness perceptions for a negatively labeled target that shows trustworthy behavior (i.e., the "trustworthy trickster") compared to a negatively labeled target that turns out to be a cheater (i.e., the "untrustworthy trickster"). The results of the two studies described in this paper confirm this reasoning and, thus, contribute to and qualify research on trustworthiness, suspiciousness, and source memory in social interactions.

Victim Sensitivity and Asymmetric Attendance to Untrustworthiness Cues

In previous studies (Suzuki and Suga, 2010; Bell et al., 2012), it has been demonstrated that people rely on cues of untrustworthiness in a person's facial appearance if they do no longer remember a person's previous behavior. Consistent with these findings, the social labels affected participants' guessing behavior in Study 1. If source memory was no longer available, participants guessed that targets associated with negative social labels had been associated with untrustworthy behavior more often than guessing that targets associated with positive social labels had been associated with trustworthy behavior. This effect was only found among participants who were classified as high in victim sensitivity and was not evident among participants low in victim sensitivity. This finding confirms and expands previous research on victim sensitivity, which showed that victim-sensitive individuals are more likely than victim-insensitive individuals to attend to social cues associated with untrustworthiness rather than trustworthiness (Gollwitzer et al., 2012). The crucial difference between this previous research and the present experiments is that previous research on asymmetrical attendance to untrustworthiness cues solely relied on self-reports (about another person's trustworthiness), whereas the first study in the present article obtained evidence for this effect in a much more unobtrusive measure: participants' guessing in the absence of source memory. This finding lends support to the "asymmetry hypothesis" formulated by Gollwitzer et al. (2013) in their "Sensitivity to Mean Intentions" (SeMI) model.

Victim Sensitivity and Asymmetrical Effects on Expectancy-Inconsistent Information

The finding that people high in victim sensitivity have a guessing bias toward untrustworthiness after being confronted with untrustworthiness-related cues is interesting in itself. However, it takes us even one step further by answering the question how people who fear to be exploited react to information that violates their initial expectation about the trustworthiness of their interaction partners.

Notably, the SeMI model does not make straightforward predictions about how victim-sensitive individuals should respond to persons who violate or confirm their initial expectations. Considering that victim-sensitive individuals experience particularly strong negative emotions when they are exploited, it seemed possible that victim sensitivity may lead to an inflexible memory advantage for cheaters. In other words, it seemed possible that victim-sensitive compared with victim-insensitive individuals recall that somebody turned out to be a cheater more easily than the information that somebody turned out to be a nice person.

In general, both types of information—the information that someone is a cheater as well as the information that someone is trustworthy—helps to decrease social uncertainty which is experienced as aversive by victim-sensitive individuals. When such information is available, remembering expectancy-inconsistent information may be particularly useful for social exchange. For instance, when people are in a situation in which cooperation is low and cheating is common, they may be extremely reluctant to cooperate with people whose previous behavior is unknown. In this situation, it is not helpful for an individual to remember particular instances of cheating because, with or without this information, this individual will refuse to cooperate (Barclay, 2008; Bell et al., 2010). Instead, it seems more functional to focus on those few cases in which the behavior is inconsistent with one's expectations about a person. Given that the effect of schema inconsistency on memory has been shown to be a fairly general phenomenon (Bell and Buchner, 2012), it seems possible that victim-sensitive individuals—due to their increased negative expectations—may show increased processing of information that specifically violates their *negative* views of other persons.

In line with this latter prediction, participants in Study 1 demonstrated better source memory for behaviors that were inconsistent with negative expectations than for behaviors that were consistent with these expectations. This finding cannot be explained by a cheater detection account, but it can be explained by an expectancy inconsistency account. Moreover, Study 1 demonstrated that this expectancy inconsistency effect was particularly evident for people high in victim sensitivity. These people, however, showed no memory advantage for behaviors

that violated positive stereotypes, which, at first glance, seems to be at odds with the SeMI model: according to this model, realizing that somebody is cheater although one expected this person to be trustworthy should be particularly aversive to victim-sensitive relative to victim-insensitive individuals. Interestingly, prior research (Bell and Buchner, 2010) found that observer-sensitive individuals (i.e., people with a true concern for the just treatment of others) do have better source memory for cheaters. In light of the present results and using an expectancy inconsistency account of memory, this prior finding might be better understood: whereas people high compared with people low in victim sensitivity have biased expectations toward targets associated with untrustworthiness, people high compared with people low in observer sensitivity might harbor more positive expectations toward targets associated with trustworthiness. Hence, learning that the scientist is untrustworthy constitutes a greater expectancy violation for the person high in observer sensitivity, whereas learning that the trickster is trustworthy constitutes a greater expectancy violation for the person high in victim sensitivity (resulting in the findings presented here and the ones observed by Bell and Buchner, 2010).

Results from Study 2 extended the findings observed on participants' source memory to participants' trustworthiness perceptions and provided experimental evidence in that regard. In this study, victim-sensitive individuals whose fear of exploitation was experimentally activated were more likely to update their trustworthiness perceptions if a negatively labeled target turned out to display trustworthy behavior. The opposite effect, that is, updating trustworthiness ratings for a positively labeled target who turned out to be a cheater, was not influenced by being victim-sensitive or victim-insensitive in Study 2. Thus, our results can be summarized as follows: victim-sensitive individuals show asymmetric expectancy violation effects which is evidenced in an asymmetric memory advantage for schema-inconsistent information as well as an asymmetric change in person perception.

Limitations

In the present studies, participants observed targets. Thus, it is not clear whether similar effects on memory and impression updating would be obtained if participants were the actual recipients of trustworthy versus untrustworthy behavior. However, research on cooperation in public goods games demonstrated that victim sensitivity is a powerful predictor of withholding contributions when cues of exploitation are present (Gollwitzer and Rothmund, 2011). Therefore, there is good reason to assume that victim sensitivity in terms of a fear of exploitation does drive expectations and expectation violations also in more interactive situations.

Another potential limitation pertains to the operationalization. Participants only judged male targets. Hence, it is unclear whether similar effects would be obtained for female targets. Moreover, the majority of participants were female. However, gender has only very small effects on victim sensitivity ($\eta^2 = 0.002$ in the validation study by Schmitt et al., 2010).

Finally, the present research demonstrates that victim-sensitive individuals react more strongly to certain types of expectancy violations. However, it is unclear which processes involved in expectancy violations drive the observed effects. Thus, whereas a purely cognitive process is possible in which victim sensitivity exacerbates contrast effects by increasing the difference in valence of the elements involved in the comparison (Biernat, 2005), it is also conceivable that victim sensitivity is related to greater feelings of surprise following expectancy evaluations. Importantly, stronger surprise alone might suffice to exacerbate contrast effects as it stimulates stronger sense-making and cognitive mastering (see Noordewier et al., 2016 for a temporal dynamics account of surprise).

CONCLUSION

In conclusion, the present research supports the hypothesis that victim sensitivity and therewith a fear of exploitation need not result in an increase in response to cheating but may ironically increase the processing of information that is inconsistent with negative stereotypes and expectations. This finding cannot be accounted for by a cheater detection explanation but nicely fits an expectancy violation account—in which victim sensitivity systematically affects participants' initial schematic expectations for dubious targets leading to stronger effects in participants' source memory and trustworthiness perceptions if these expectations are not met. This might in fact be interpreted as good news, as it suggests that even persons with a high fear of exploitation are able to overcome their habitual negative expectations toward their social environment when they are confronted with more valid information about another person's trustworthiness.

ETHICS STATEMENT

The reported studies were exempt from requiring an approval of an ethics committee. Studies were voluntary, anonymous, non-invasive, and did not involve deception. Approval of an ethics committee for such studies is not an requirement in Germany where the studies were conducted. Participants received general information about the study before taking part, were informed that participation is voluntary, and that participation can be withdrawn.

AUTHOR CONTRIBUTIONS

PS drafted the manuscript. MG, LM, AB, and RB revised the manuscript. Study 1 was designed and analyzed by LM, AB, and RB. Study 2 was designed and analyzed by PS and MG.

REFERENCES

Barclay, P. (2008). Enhanced recognition of defectors depends on their rarity. *Cognition* 107, 817–828. doi: 10.1016/j.cognition.2007.11.013

Barclay, P., and Lalumière, M. (2006). Do people differentially remember cheaters? *Hum. Nat.* 17, 98–113. doi: 10.1007/s12110-006-1022-y

Bayen, U. J., and Kuhlmann, B. G. (2011). Influences of source-item contingency and schematic knowledge on source monitoring: tests of the probability-matching account. *J. Mem. Lang.* 64, 1–17. doi: 10.1016/j.jml.2010.09.001

Bayen, U. J., Murnane, K., and Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 197–215. doi: 10.3758/PBR.17.4.465

Bell, R., and Buchner, A. (2010). Justice sensitivity and source memory for cheaters. *J. Res. Pers.* 44, 677–683. doi: 10.1016/j.jrp.2010.08.011

Bell, R., and Buchner, A. (2012). How adaptive is memory for cheaters? *Curr. Dir. Psychol. Sci.* 21, 403–408. doi: 10.1177/0963721412458525

Bell, R., Buchner, A., Kroneisen, M., and Giang, T. (2012). On the flexibility of social source memory: a test of the emotional incongruity hypothesis. *J. Exp. Psychol.* 38, 1512–1529. doi: 10.1037/a0028219

Bell, R., Buchner, A., and Musch, J. (2010). Enhanced old-new recognition and source memory for faces of cooperators and defectors in a social-dilemma game. *Cognition* 117, 261–275. doi: 10.1016/j.cognition.2010.08.020

Bell, R., Mieth, L., and Buchner, A. (2015). Appearance-based first impressions and person memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 456–472. doi: 10.1037/xlm0000034

Biernat, M. (2005). *Standards and Expectancies – Contrast and Assimilation in Judgments of Self and Others*. New York, NY: Psychology Press.

Buchner, A., Bell, R., Mehl, B., and Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evol. Hum. Behav.* 30, 212–224. doi: 10.1016/j.evolhumbehav.2009.01.004

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31, 187–276. doi: 10.1016/0010-0277(89)90023-1

Cosmides, L., and Tooby, J. (1989). Evolutionary psychology and the generation of culture: II. Case study: a computational theory of social exchange. *Ethol. Sociobiol.* 10, 51–97. doi: 10.1016/0162-3095(89)90013-7

Cosmides, L., and Tooby, J. (2005). "Neurocognitive adaptations designed for social exchange," in *The Handbook of Evolutionary Psychology*, ed. D. M. Buss (Hoboken, NJ: John Wiley & Sons Inc), 584–627.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., and Nadarevic, L. (2009). Multinomial processing tree models. A review of the literature. *J. Psychol.* 217, 108–124.

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Gollwitzer, M., and Rothmund, T. (2009). "When the need to trust results in unethical behavior: the sensitivity to mean intentions (SeMI) model," in *Psychological Perspectives on Ethical Behavior and Decision Making*, ed. D. De Cremer (Charlotte, NC: Information Age), 135–152.

Gollwitzer, M., and Rothmund, T. (2011). What exactly are victim-sensitive persons sensitive to? *J. Res. Pers.* 45, 448–455. doi: 10.1016/j.jrp.2011.05.003

Gollwitzer, M., Rothmund, T., Alt, B., and Jekel, M. (2012). Victim sensitivity and the accuracy of social perceptions. *Pers. Soc. Psychol. Bull.* 38, 975–984. doi: 10.1177/0146167212440887

Gollwitzer, M., Rothmund, T., Pfeiffer, A., and Ensenbach, C. (2009). Why and when justice sensitivity leads to pro- and antisocial behavior. *J. Res. Pers.* 43, 999–1005. doi: 10.1016/j.jrp.2009.07.003

Gollwitzer, M., Rothmund, T., and Süssenbach, P. (2013). The sensitivity to mean intentions (SeMI) model: basic assumptions, recent findings, and potential avenues for future research. *Soc. Psychol. Pers. Compass* 7, 415–426. doi: 10.1111/spc3.12041

Gollwitzer, M., Schmitt, M., Schalke, R., Maes, J., and Baer, A. (2005). Asymmetrical effects of justice sensitivity perspectives on prosocial and antisocial behavior. *Soc. Just. Res.* 18, 183–201. doi: 10.1007/s11211-005-7368-1

Graesser, A. C., and Nakamura, G. V. (1982). "The impact of a schema on comprehension and memory," in *The Psychology of Leaning and Motivation*, ed. G. H. Bower (New York: Academic Press), 59–109.

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis*. New York, NY: The Guilford Press.

Mealey, L., Daood, C., and Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethol. Sociobiol.* 17, 119–128. doi: 10.1016/0162-3095(95)00131-X

Mieth, L., Bell, R., and Buchner, A. (2016). Memory and disgust: effects of appearance-congruent and appearance-incongruent information on source memory for food. *Memory* 24, 629–639. doi: 10.1080/09658211.2015.1034139

Moshagen, M. (2010). multiTree: a computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* 42, 42–54. doi: 10.3758/BRM.42.1.42

Noordewier, M. K., Topolinski, S., and Van Dijk, E. (2016). The temporal dynamics of surprise. *Soc. Pers. Psychol. Compass* 10, 136–149. doi: 10.1111/spc3.12242

Oda, R. (1997). Biased face recognition in the prisoner's dilemma game. *Evol. Hum. Behav.* 18, 309–315. doi: 10.1016/S1090-5138(97)00014-7

Phillips, P. J., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image Vision Comput.* 16, 295–306. doi: 10.1016/S0262-8856(97)00070-X

Schmitt, M., Baumert, A., Gollwitzer, M., and Maes, J. (2010). The justice sensitivity inventory: factorial validity, location in the personality facet space, demographic pattern, and normative data. *Soc. Justice Res.* 23, 211–238. doi: 10.1007/s11211-010-0115-2

Schmitt, M., Gollwitzer, M., Maes, J., and Arbach, D. (2005). Justice sensitivity: Assessment and location in the personality space. *Eur. J. Psychol. Assess.* 21, 202–211. doi: 10.1027/1015-5759.21.3.202

Suzuki, A., and Suga, S. (2010). Enhanced memory for the wolf in sheep's clothing: facial trustworthiness modulates face-trait associative memory. *Cognition* 117, 224–229. doi: 10.1016/j.cognition.2010.08.004

Todorov, A., Pakrashi, M., and Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Soc. Cogn.* 27, 813–833. doi: 10.1521/soco.2009.27.6.813

Yamagishi, T. (2001). "Trust as a form of social intelligence," in *Trust in Society*, Vol. 2, ed. K. S. Cook (New York, NY: Russell Sage Foundation), 121–147.

Zebrowitz, L. A., and Montepare, J. M. (2008). Social psychological face perception: why appearance matters. *Soc. Pers. Psychol. Compass* 2, 1497–1517. doi: 10.1111/j.1751-9004.2008.00109.x

Conflict of Interest Statement: 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.

Copyright © 2016 Süssenbach, Gollwitzer, Mieth, Buchner and Bell. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 October 2016
doi: 10.3389/fpsyg.2016.01624

Functional Significance of Conflicting Age and Wealth Cross-Categorization: The Dominant Role of Categories That Violate Stereotypical Expectations

Jingjing Song and Bin Zuo *

School of Psychology, Central China Normal University, Wuhan, China

The purpose of the current study was to identify the functional significance of conflicting stereotypes and to identify the dominant category in such conflicts. In the present research we examined the conflicting crossed categories of age and wealth with regard to warmth and competence perceptions. It was found (Pilot Study and Study 1) that the old-rich targets presented a conflicting stereotype group in the perception of warmth, whereas young-poor targets presented a conflicting stereotype group in the perception of competence. In addition, the *old* stereotype dominated the warmth evaluation of old-rich targets, whereas the *poor* stereotype dominated the competence evaluation of young-poor targets. In Study 2, participants provided warmth and competence evaluations after they learned about the targets' behaviors which demonstrated high or low warmth and high or low competence. The results suggest that for the warmth evaluation of the old-rich target the category that did not match the behavior (i.e., contradicted the stereotype expectation) was more salient and drove judgments. However, the effect of stereotype expectation violation was not found in the competence evaluation of the young-poor target. The results are discussed in terms of their implications for understanding factors that activate and inhibit stereotyped perceptions.

Keywords: stereotype, cross-categorization, age, wealth, functional significance

OPEN ACCESS

Edited by:
Mario Gollwitzer,
University of Marburg, Germany

Reviewed by:
Philipp Süssenbach,
University of Marburg, Germany
Daniela Bernhardt,
University of Erlangen-Nuremberg, Germany

***Correspondence:**
Bin Zuo
zuobin@mail.ccnu.edu.cn

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *25 April 2016*
Accepted: *04 October 2016*
Published: *21 October 2016*

Citation:
Song J and Zuo B (2016) Functional Significance of Conflicting Age and Wealth Cross-Categorization: The Dominant Role of Categories That Violate Stereotypical Expectations. Front. Psychol. 7:1624. doi: 10.3389/fpsyg.2016.01624

INTRODUCTION

People live in complicated societies and may belong to many social categories simultaneously. Importantly, how a person is perceived may vary depending on those categories or combinations of categories. Cross-categorization refers to the process of classifying persons according to two categories. Recently, research has begun to explore people's evaluations of persons belonging to crossed-categories (Urada et al., 2007; Bodenhausen, 2010; Sesko and Biernat, 2010; Freeman and Ambady, 2011; Johnson et al., 2012; Neuberg and Sng, 2013; Penner and Saperstein, 2013; Kang and Bodenhausen, 2015). The accessibility and functional significance of each category can affect perceptions of the cross-categorized target. The functional significance of a category refers to the dominance of the category's influence on the perceiver's evaluations of a target in a specific situation It has been demonstrated that the perceiver's attributes and the context can influence the functional significance of a category when considered in isolation. Extending previous research

(Turner et al., 1987; Crisp and Hewstone, 2006), in the present study we focused on the age and wealth crossed-category and tested the co-effect of the perceiver's attributes (attitudes toward each simple-category) and the context (the behavior of the target) on functional significance.

The Conflicting Natural and Social Crossed Categories

Previous research on the perception of cross-categorized targets focused on explicit and relatively unmodifiable natural categories, such as gender or race, including the cross of race and gender (Johnson et al., 2012; Klauer et al., 2014; Schug et al., 2015), race and age (Kang et al., 2014), and age and gender (Klauer et al., 2003; Cloutier et al., 2014). One exception was a study by Smith et al. (1996), in which both categories were social categories (i.e., a target could be categorized as a baseball player and a gambler as well).

Major natural categories dominate the early stages of person perception (Dovidio et al., 1997; Ma and Correll, 2011). However, when the perceiver has sufficient time, capacity, and motivation, targets will be categorized in terms of multiple categories (Pendry and Macrae, 1996), which include natural categories as well as social categories. In this context social categories may play an important role in stereotype evaluation. One social category that perceivers commonly use is Social Economic Status, and based on the wealth component of this category, individuals can be classified as rich and poor targets. This category was one focus of the current study.

Stereotypes about the rich and poor can be conceptualized in terms of Fiske et al. (2002) stereotype content model (SCM), which assumes that people tend to evaluate others based on the two dimensions of warmth and competence. Competence refers to position in the status power hierarchy, whereas warmth refers to cooperation within one's own group (Abele and Wojciszke, 2014). The stereotype link to the rich is low warmth and high competence (Piff et al., 2010). However, the poor target is perceived as having low warmth and low competence (Fiske et al., 2002). The negative stereotype of the poor may decrease their probability of receiving an equal professional development opportunity, and perceptions of an imbalance in social wealth distribution may lead to hatred of the rich, thus further increasing the risk of social instability. Therefore, research on the wealth category is of great applied value.

With regard to the natural category examined in this study, we focused on age, which has received less attention than other major natural categories such as sex or race. The old target is perceived as showing high warmth (Kite et al., 2005; Chasteen et al., 2012) and low competence (Hess et al., 2009; Eich et al., 2014), whereas the young target is perceived as showing low warmth and high competence (Song et al., in press).

The old-rich target elicits conflicting stereotypes, as there is a high warmth and low competence evaluation for *old*, but a low warmth and high competence evaluation for *rich*. The young-poor target is also in a conflicting stereotype group as there is a high competence evaluation of the young target and low competence evaluation of the poor target. In cross-categorization involving conflicting stereotypes, the two sub-categories, which refer to *old* and *rich* categories when evaluating the old-rich target, and *young* and *poor* categories when evaluating the young-poor target, are subject to a competition for mental dominance, and they may not have equal psychological significance to the perceiver (Crisp and Hewstone, 2006). Thus, the salient and dominant category in conflicting crossed-categories (i.e., the category with high functional significance) can determine the perception of this target as negative or positive. In the current study, we only focused on this aspect of cross-categorized groups, namely the functional significance of conflicting stereotypes.

The Relation between the Perceiver's Attributes and Functional Significance

In the case of cross-categorization, the most relevant and accessible category will stand out (Bodenhausen, 2010), and when context information is not given, the degree of accessibility is determined by two factors about the perceiver's attributes. First, accessibility is determined by the strength of the perceiver's attitude (stereotype) toward the relevant categories. The category about which people have strong attitudes tends to attract attention and to be the dominant category (Fazio et al., 1995; Crisp and Hewstone, 2006). Second, accessibility is determined by the perceiver's past experience in categorizing a particular person or other social object. The perceiver who had previously judged the target as belonging to one category would likely categorize the target in a similar way in the future (Smith et al., 1992).

SCM predicts conflicting stereotypes of old-rich targets with respect to both warmth and competence, but for young-poor targets, the conflicting stereotype occurs only with respect to competence. Thus, we focused on the functional significance of the categories *old* and *rich* with respect to evaluations of both warmth and competence, but for the categories of *young* and *poor* we focused only on the evaluation of competence. The stereotype strength is the key factor influencing the dominant category (Fazio et al., 1995; Crisp and Hewstone, 2006). However, the warmth stereotype strength is similar for old and rich targets (Fiske et al., 2002). Thus, it is difficult to recognize the dominant category of the old-rich target in the warmth evaluation only based on the stereotype strength. Furthermore, as a fundamental natural category, age is more visible and easily identifiable than wealth, and thus it is more accessible. Therefore, age is repeatedly used to categorize people in daily life, and this repeated practice may make the age category more accessible to the perceiver than the wealth category. Thus, we assumed that in the warmth evaluation of the conflicting categorization (old-rich), the age category would be the dominant category (H1).

For the competence evaluation of the old-rich and young-poor targets, the age category is also more accessible. However, the perception of competence is always closely connected with a person's wealth (Cuddy et al., 2009), and the stereotypes of competence in relation to the wealth category were shown to be stronger than the stereotypes of competence in relation to the age category in a previous investigation (Fiske et al., 2002). Thus, the competence evaluation results in an emphasis on the

wealth category. We assumed that in the competence evaluation of conflicting categories (old-rich and young-poor), the wealth category would be the dominant category (H2).

The Relation between Context Information and Functional Significance

Context information and specific mental schemas (stereotypes) that are triggered by a combination of categories can influence perception of the crossed-category target (Casper et al., 2011). The parallel constraint satisfaction model suggests that different context information activates different subsets of a network of connections, and thus the dominant category and the evaluation of the target would change across contexts (Kunda et al., 1997; Crisp and Hewstone, 2006). One important question is how context moderates perceptions of crossed-category targets, to understand what circumstances give rise to activation of some stereotype components and inhibition of others.

One form of context can be situational factors, e.g., a white female target in a group of black and white men (Van Rijswijk and Ellemers, 2002), and it has been demonstrated that the category that is unique, clear, and prominent in the situation is the dominant category. Context can also be the behavior of the target, and it was this kind of context that was examined in the current study on the age and wealth cross-categorization. Specifically, we analyzed the co-effect of (a) the perceiver's attitude (stereotype) toward the relevant categories and (b) context in the form of the target's specific behaviors, on the functional significance of each category.

Turner et al. (1987) showed that, when the behavior is consistent with the stereotype of a particular category (nominative fit), this category would be the dominant category, as the behavior information directs the perceiver to the stereotype-consistent category. However, a large number of studies since then have found that the perceiver pays attention to the target that contradicts the stereotype expectancy (Bettencourt et al., 1997; Dickter and Gyurovski, 2012; Garcia-Marques et al., 2016; Jerónimo et al., 2016). People engage in more effortful cognitive processing (Jerónimo et al., 2016), reorganize the "wrong description" (contradicting the expectancy), perceive the stereotype-inconsistent target as atypical, make more explanations about and prefer external attributions for the behavior (Sekaquaptewa et al., 2003; Sekaquaptewa and Espinoza, 2004), and make the stereotype-inconsistent behavior conform to their stereotype. Furthermore, the shifting standards model suggests that the target is evaluated with reference to the stereotype expectations of that particular category; the perceiver makes extreme judgments (Biernat and Vescio, 2002) and uses ironic language more often in response to the target who behaves in a manner contrary to expectation (Burgers and Beukeboom, 2014), so as to maintain stereotypic expectancies. Thus, the category associated with counter-stereotypic behavior attracts more attention, and we suspected that these categories may dominate the perception of cross-categorized targets.

In the warmth evaluation of the conflicting categorization of old-rich targets, we assumed that counter-stereotypic behavior would dominate the perception of cross-categorized targets. When the old-rich target shows low warmth behavior, the low warmth behavior contradicts the stereotype expectation of someone who is old. Thus, the *old* category would be expected to be the dominant categorization in the perception of the old-rich target that shows low warmth behavior (H3a). By contrast, the *rich* categorization would be the dominant categorization in the perception of the old-rich target that shows high warmth behavior (H3b). Furthermore, in the competence evaluation of the conflicting cross-categorized target (i.e., young-poor and old-rich), when the target fails, the *young* categorization should be the dominant categorization for the young-poor target, and the *rich* categorization should be the dominant categorization for the old-rich target (H3c); when the target succeeds, the *poor* categorization should be the dominant categorization for the young-poor target, and the *old* categorization should be the dominant categorization for the old-rich target (H3d).

Overview of the Current Study

We conducted a pilot study and two larger studies to test the functional significance of conflicting stereotypes, focusing on the cross-categorization of age and wealth. The purpose of the pilot study was to verify that these categories did generate conflicting stereotypes, and furthermore to evaluate the functional significance of each category in cross-categorized groups. A categorization task was used in which the participants categorized the target (using both simple categories and crossed categories) as showing high competence or low competence and as showing high warmth or low warmth. In Study 1, the participants used rating scales to evaluate the warmth and competence of the young, the old, the rich, the poor, the old-rich and the young-poor target. Regression analysis was conducted to test the effect of attitude about the simple-category target on the perception of the cross-categorized target, and relative weight analysis was conducted to quantitatively analyze the functional significance of each category. Study 2 tested the functional significance in specific scenarios to determine whether the dominant and weaker categories varied depending on context, and tested the co-effect of the "attitude about the simple-category" and the "behavior of the target" on the perception of the cross-categorization. The dependent measures included direct (warmth and competence evaluations) and indirect (attributions) stereotype evaluations.

PILOT STUDY: THE EXAMINATION OF AGE AND WEALTH CATEGORIES IN TERMS OF THE SCM AND QUALITATIVE ANALYSIS OF FUNCTIONAL SIGNIFICANCE

The first purpose of the pilot study was to verify the SCM in relation to the age and wealth categories. We expected to find that *old* would be the high warmth and low competence group, *young* and *rich* would be the high competence and low warmth groups, and *poor* would be the low warmth and low competence group, constituting conflicting stereotype groups. The second purpose of the pilot study was to qualitatively explore the functional significance of the conflicting stereotypes.

Method

Participants

A total of 45 students from a university in central China volunteered to participate in this study. Data from one additional participant were not included in the analysis because of an incomplete categorization task. The participants' ages ranged from 18 to 30 years ($M = 22.78$, $SD = 3.00$), and there were 18 males (40.0%) and 27 females (60.0%).

Materials and Procedure

Permission was obtained from the university ethics committees. Participants were recruited and invited to the laboratory where they were introduced to the categorization task. After receiving a description of the study, participants gave written informed consent. They received a small gift (candy) at the end of the study.

The Categorization Task: The task materials consisted of eight "identity cards." Each identity card presented a very simple head-and-shoulders photo silhouette in black-and-white on the left side of the card (Dommelen et al., 2015) and the identity information in text on the right side of the card. Samples of identity cards can be seen in the Supplementary Material. The participants could not recognize the gender or age through the photos, and the identity text was the only useful information. On four identity cards, a simple category was depicted (*old, young, poor,* and *rich*). Crossing wealth and age led to four category conjunctions, which were presented on the other four cards (the old-rich, the young-rich, the old-poor, and the young-poor). Each identity card was presented three times to each participant.

The first task was the warmth categorization, in which participants placed targets into one of two boxes labeled "high warmth (kindness and friendliness)" and "low warmth (kindness and friendliness)." The second task was the competence categorization, in which participants placed targets into one of two boxes labeled "high competence (confidence and intelligence)" and "low competence (confidence and intelligence)." After being shown an example identity card, participants were asked to categorize the full set of cards into the two boxes (high vs. low warmth and high vs. low competence). The cards were presented in random order, and the participants were given enough time to evaluate the targets and finish the task. Participants were de-briefed after the task.

The frequency with which each participant assigned a specific card into the high competence box or the high warmth box was calculated, and it ranged from 0 to 3 because each card was presented three times. The categorization task has been widely used in previous research by assigning the target to the "in-group (us)" box or the "out-group (not us)" box, and it has been demonstrated to be an effective task to categorize the cross-categorized target (Singh et al., 1997; Dommelen et al., 2015).

Results

Chi-square tests were conducted to compare the number of people choosing the high competence/warmth box for each target 0, 1, 2, or 3 times with the expected value of 45/4 = 11.25. As shown in **Table 1**, the results showed that 23 subjects assigned all three *old* targets into the high warmth box, which is significantly higher than the expected value ($\chi^2 = 21.76$, $p < 0.001$). The frequencies with which participants placed all three young targets into the high competence ($n = 29$) and high warmth boxes ($n = 20$) were also significantly higher than the expected value ($\chi^2 = 47.36$, $p < 0.001$; $\chi^2 = 10.20$, $p < 0.05$). As for wealth, the frequencies with which subjects placed all three rich targets into the high competence ($n = 34$) and high warmth ($n = 3$) boxes and assigned all three poor targets into the low competence ($n = 31$) box were both significantly different compared with the expected value of 11.25 ($\chi^2 = 65.13$, $p < 0.001$; $\chi^2 = 12.\ 51$, $p < 0.01$; $\chi^2 = 50.20$, $p < 0.001$). The results indicate that the young-poor target presents a conflicting stereotype group in the competence evaluation, and the old-rich target presents a conflicting stereotype group only in the warmth evaluation.

TABLE 1 | Absolute frequency of categorizations for the high warmth and the high competence category in the pilot study.

	High warmth category				High competence category			
	0	**1**	**2**	**3**	**0**	**1**	**2**	**3**
Old	2	7	13	23	12	13	8	12
Rich	14	19	9	3	1	1	9	34
Young	6	8	11	20	1	1	14	29
Poor	7	10	18	10	31	10	3	1
Old-rich	3	8	12	22	1	3	6	35
Young-poor	12	10	17	6	21	5	11	8
Old-poor	7	13	11	14	38	5	1	1
Young-rich	16	9	14	6	5	3	7	30

N = 45, the data in the table refers to the number of subjects assigned targets 0, 1, 2, or 3 times into the high warmth/competence category. As young-poor target presents a conflicting stereotype group in the competence evaluation (young target is perceived as high competence, and poor target is perceived as low competence), and great majority of subjects assigned all three young-poor targets into the low competence category. Thus, we could assume the strength of the high competence stereotype of the young target was lower than the strength of the low competence stereotype of the poor target, and the poor was the primary category in the competence evaluation of the young-poor target.

Furthermore, the frequencies with which participants placed all three old-rich targets into the high warmth ($n = 22$) box and high competence ($n = 35$) box, and assigned all three young-poor targets into the low competence ($n = 21$) box were significantly higher than the expected value of 11.25 ($\chi^2 = 17.31$, $p < 0.01$; $\chi^2 = 67.98$, $p < 0.001$; $\chi^2 = 12.87$, $p < 0.01$). Thus, for warmth evaluations of the old-rich target, *old* was the primary category, and in the competence evaluation of the young-poor target, *poor* was the primary category. H1 was supported. H2 originally made predictions about both the competence evaluation of the young-poor target and of the old-rich target, but we were only able to test the former component because in our pilot study *old* and *rich* did not constitute a conflicting cross-category in the competence evaluation. Thus, the dominance of the wealth category could not be fully confirmed. The results regarding the young-rich and old-poor combinations can be seen in Appendix A (Supplementary Material).

Discussion of Pilot Study

The results of the pilot study do not fully comply with the SCM, as old targets were not perceived as incompetent. One possible methodological reason is that the term "old" used in the stimulus material may have been interpreted differently by the participants than was intended, and there may be large differences especially concerning competence if one thinks of an "old" person who is 55 or 90 years old. Another reason may be that part of the definition of competence, namely intelligence, could have been interpreted as meaning either fluid intelligence or crystallized intelligence, or both. Furthermore, young targets were perceived as warm rather than cold. This may be because the participants were in-group members of the *young* category, and thus they have made positive evaluations of the young target.

Although not fully consistent with the SCM, the pilot data do provide two examples of conflicting stereotype groups (old-rich target in the warmth evaluation and young-poor target in the competence evaluation), and these could be used as stimuli in Studies 1 and 2. Furthermore, the results showed that, *old* was the dominant category in the warmth evaluations of the old-rich target, and *poor* was the dominant category in the competence evaluation of the young-poor target.

There were limitations in the identity card categorization task used in the pilot study. Participants could only classify targets as being in the "high" or "low" warmth/competence-group, and there was no "middle" or "cannot decide" category. This may lead to an overestimation of stereotypical trait ascriptions, as subjects were forced to choose either the "high" or the "low" box even if they had no clear preference. In addition, using each identity card three times to represent the distinct categories and category conjunctions artificially increases power. Furthermore, demand characteristics may have played a role, and some participants might have been aware of what was being measured by the task.

The pilot study identified the dominant category in the warmth evaluation of the old-rich target, and in the competence evaluation of the young-poor target. However, the pilot study did not assess the relative weight of each category quantitatively. Thus, quantitative analyses were needed to directly compare the relative weight of each category.

STUDY 1: THE RELATIVE WEIGHT OF SIMPLE CATEGORIES IN THE PERCEPTION OF CONFLICTING STEREOTYPES

The pilot study demonstrated the functional significance of each of the simple categories to which the old-rich and young-poor targets belonged. However, the pilot study did not assess stereotype quantitatively, and it did not calculate the relative weight of each category in the stereotype evaluations of the cross-categorized targets. In Study 1, we used regression analysis to directly examine the relative importance (weight) of each category by asking the participants to use rating scales to evaluate the warmth and competence of targets belonging to simple categories and crossed-categories.

Methods

Participants

A total of 104 students from a university in central China participated in this study. The participants' ages ranged from 17 to 23 years ($M = 19.38$, $SD = 1.17$), and there were 20 males (19.2%) and 84 females (80.8%). There were 51 participants from rural areas (49.0%) and 53 from the city (51.0%). When asked to rate how wealthy they were, three participants described themselves as "very poor" (2.9%), 29 participants described themselves as "rather poor" (27.9%), 62 participants described themselves as "average" (59.6%), and 10 participants described themselves as "rather rich" (9.6%).

Materials and Procedure

Permission was obtained from the university ethics committees to conduct this study. The questionnaire was administered to the students in a class during one class period. Two trained data collectors administered the questionnaire according to a manual of procedures to standardize the data collection process. Participants gave written informed consent after receiving a description of the study, and they received a small gift (candy) at the end of the study.

The participants were asked to evaluate the warmth and competence of six targets (the old, the rich, and the old-rich targets; the young, the poor, and the young-poor targets). As an introduction, participants were told that this was a social perception task, and they were asked to evaluate some strangers' personalities on the basis of a limited amount of information. The six identity targets were presented on identity cards (refer to pilot study). The presentation order of the targets was counterbalanced, and it complied with the principle that the first two cards presented were simple-category cards (e.g., old, rich), followed by a card crossing the two simple categories (e.g., old-rich). Competence was evaluated with three traits: competence, intelligence, and confidence. Warmth was also evaluated with three traits: warmth, friendliness, and kindness (Fiske et al., 2002; Judd et al., 2005). Participants were asked to rate each adjective according to its descriptiveness of the target on a scale from 1 (*not at all descriptive*) to 5 (*very descriptive*). The sum of the three items (warmth or competence) was the final score. The higher the score was, the higher the perceived competence or warmth of the target was. The participants were then de-briefed. This method, measuring explicit attitudes toward the target, has been widely used in previous research, and it has been shown to be valid (Judd et al., 2005; Corcoran et al., 2009; Kang et al., 2014). In the current study, the internal consistency reliability (α) was 0.76 for the competence measure and 0.86 for the warmth measure.

Results

The correlations presented in **Table 2** indicated that the warmth evaluation of the old-rich target was significantly positively correlated with the warmth evaluation of the rich target and of the old target. Additionally, the competence evaluation of the young-poor target was significantly positively correlated with the competence evaluation of the poor target, although not correlated with the competence evaluation of the young target. The results about the correlation among the competence ratings

TABLE 2 | Correlations among all the variables in study 1 and descriptive information.

	1. Old warmth	2. Rich warmth	3. Old-rich warmth	4. Young competence	5. Poor competence	6. Young-poor competence
1	–					
2	0.28**	–				
3	0.60**	0.49**	–			
4	0.31**	0.32**	0.37**	–		
5	0.17	0.23*	0.15	0.22*	–	
6	0.02	−0.05	0.20*	0.12	0.35**	–
M	10.79	8.00	10.45	10.85	8.34	8.71
SD	2.15	2.20	2.27	1.82	2.17	2.20

*N = 104, *p < 0.05, **p < 0.01, scale range: 3–15.*

for old, rich, and old-rich target, and the warmth ratings for young, poor, and young-poor targets can be seen in Appendix B (Supplementary Material).

We next used linear regressions to explore the relationship between the evaluation of targets belonging to the simple categories and crossed categories. For the warmth evaluation of the old-rich target, we conducted a regression analysis to test the warmth evaluations of the old and rich targets as predictors of the warmth evaluation of the old-rich target. The first block included the demographic variables: age, gender, wealth, and Hukou (a family registration program that serves as a domestic passport and divides residents into two groups: urban and rural). The second block included the warmth evaluation of the rich and the old targets. The warmth evaluation of the old-rich target was the dependent variable. As can be seen in **Table 3**, the regression analysis indicated that, after controlling for the demographic variables, the warmth evaluations of the old target and the rich target were significantly associated with the warmth evaluation of the old-rich target ($\beta = 0.51, p < 0.001; \beta = 0.34, p < 0.001$). Next we tested if one of the simple categories was dominant. Relative weight (*RW*) analysis is a useful technique to calculate the relative importance of predictors (independent variables) when they are correlated with each other (LeBreton and Tonidandel, 2008). This analysis indicated that the relative weight of the *old* category ($RW = 0.30$) was greater than the relative weight of the *rich* category ($RW = 0.17$), providing further support for H1.

In the competence evaluation of the young-poor target, we also conducted a regression to analyze the competence evaluations of the young and poor targets as predictors of the competence evaluation of the young-poor target. As can be seen in **Table 3**, the results indicated that, after controlling for the demographic variables, the competence evaluation of the poor target was a significant predictor of the competence evaluation of the young-poor target ($\beta = 0.35$, $p < 0.01$), but the competence evaluation of the young target was not a significant predictor ($\beta = 0.04$, $p > 0.05$). The relative weight of the *poor* category ($RW = 0.11$) was greater than the relative weight of the *young* category ($RW = 0.01$), providing further support for H2 in the competence evaluation of the young-poor target. However, the competence evaluation of old-rich target was not tested in the current paper as the old-rich target was not a conflicting cross-category in the competence evaluation, and so the dominance of the wealth category in the competence evaluation of the old-rich target could not be confirmed. The results on the competence ratings of the old, rich, and old-rich, and warmth ratings of the young, poor, and young-poor can be seen in Appendix C (Supplementary Material).

Discussion of Study 1

The results of Study 1 were consistent with the results of the pilot study, further documenting that in the warmth evaluation of the old-rich target, the *old* category was the dominant category, and in the competence evaluation of the young-poor target, the *poor* category was the dominant category. Moreover, the results of Study 1 indicated that the stereotype evaluation of the simple category was positively related to the stereotype evaluation of the crossed category, and the strength of the perceiver's stereotype of each category determined the dominant category in the perception of the crossed-category target. In addition, as the context plays an important role in the functional significance of conflicting stereotypes, in Study 2 we tested the functional significance in specific scenarios.

STUDY 2: THE SCENARIO SPECIFICITY OF FUNCTIONAL SIGNIFICANCE

Study 2 tested the functional significance of conflicting stereotypes in specific scenarios. Participants were asked to use rating-scales to provide a direct index of their evaluations of the warmth or competence of targets with different behaviors. Moreover, inspired by the stereotype explanatory bias approach (Sekaquaptewa et al., 2003; Sekaquaptewa and Espinoza, 2004), the participants were asked to make attributions about the behavior of the target, and the attributions were taken as an indirect index of their perceptions of warmth or competence.

Methods

Participants

A total of 156 students from a university in central China participated in this study. Of these participants, 95 evaluated the warmth of the old, the rich, and the old-rich targets, and 61 evaluated the competence of the young, the poor, and the young-poor targets. The participants' ages ranged from 17 to 27 years ($M = 19.81$, $SD = 1.66$), and there were 41 males (26.3%) and 115 females (73.7%). There were 69 participants from rural areas (44.2%) and 87 from the city (55.8%). When asked to rate how wealthy they were, 5 participants described themselves as "very poor" (3.2%), 36 participants described themselves as "rather poor" (23.1%), 110 participants described themselves as "average" (70.5%), and 5 participants described themselves as "rather rich" (3.2%).

Materials and Procedure

Permission was obtained from the university ethics committees to conduct this study. Participants volunteered to participate for extra course credit. They provided informed consent and were de-briefed after the study. They received a small gift (candy) for their participation.

TABLE 3 | Hierarchical Linear Models of simple-category evaluations in relation to crossed-category evaluations ($N = 104$).

	Warmth evaluation				Competence evaluation		
	β	***t***	***RW***		***β***	***t***	***RW***
Age	0.10	0.12		Age	0.10	0.93	
Gender	0.03	0.32		Gender	−0.05	−0.50	
Wealth	0.07	0.87		Wealth	0.02	0.15	
Hukou	−0.03	−0.38		Hukou	−0.17	−1.76	
Old	0.51	6.44***	0.30 (62.84%)	Young	0.04	0.42	0.01 (6.65%)
Rich	0.34	4.39***	0.17 (37.16%)	Poor	0.35	3.56**	0.11 (93.35%)
	$R^2 = 0.48$				$R^2 = 0.17$		

*RW = raw relative weight, and numbers in brackets refer to rescaled relative weight estimates reported as percentage of predicted variance. Hukou is a household registration system in China, and it includes two types: rural and city, 1 = city, 2 = rural. For Gender, 1 = male, 2 = female; **p < 0.01, ***p < 0.001.*

Six psychology doctoral students screened and chose four scenarios (high competence, low competence, high warmth, and low warmth) from the Judd et al. (2005) list of scenarios, and the scenario nominated most in each category was the final scenario used for that category. The following scenarios were chosen: one high warmth scenario (*X helped a blind woman cross the street*), one low warmth scenario (*X could not be bothered to give directions to a stranger*), one high competence scenario (*X won the yearly award for the employee who contributes most to the company's profits*), and one low competence scenario (*X failed a job interview*).

The introduction was the same as in Study 1. The description of the stranger added information about context in the form of behavior information. In order to minimize the demand characteristics and prevent the participants from guessing the purpose of the study, participants were randomly assigned to finish one of two tasks: (a) evaluate the high or low warmth behavior of the old, rich, and old-rich targets, or (b) evaluate the high and low competence behavior of the young, poor, and young-poor targets. The presentation order of the two scenarios (high and low competence or warmth) was random. For each scenario, the participants were firstly asked to evaluate the warmth or competence of two simple-category targets (e.g., *X was an old or rich person, and X helped a blind woman cross the street*), and then asked to evaluate the warmth or competence of the crossed-category target (e.g., *X was an old-rich person, and X helped a blind woman cross the street*). The warmth and competence rating scales were the same as in Study 1. In the current study, warmth and competence evaluation measures both had good reliability ($\alpha = 0.86$, $\alpha = 0.78$).

Moreover, after the direct warmth or competence evaluation, the participants were told to think carefully about why the stranger was engaging in the behavior based on the limited amount of information provided, and they were asked to write down one plausible explanation. The participants were being asked to make an attribution about the high or low competence behavior of the young, poor, and young-poor targets or about the high or low warmth behavior of the old, rich, and old-rich targets. Participants were de-briefed after the study.

Coding of Attributions

In the current study, each attribution was rated by the research team on a five-point scale based on the attribution positivity. A positive attribution means that in the participant's view, the target is showing high warmth or high competence. Coding was conducted in three steps: (1) creation of an attribution table, (2) creation of a coding manual, and (3) conversion. In the first step, two experts in the stereotype field reviewed all attributions provided by all the participants, and these attributions were classified based on shared semantic meaning. They discussed any disagreements and compiled an attribution table based on consensus. The categories of the attributions in each scenario can be seen in Appendix D (Supplementary Material).

In the second step, creation of a coding manual, six doctoral students in the stereotype field rated the positivity of each category of attributions summarized in the first step. Five points were used, ranging from 1 (*in the participant's view, X was not at all warm or competent*) to 5 (*in the participant's view, X was very warm or competent*). The higher the score was, the higher the attribution positivity was. The integer of the average of the six raters was the final "attribution positivity" score for that type of attribution. For example, in the warmth evaluation of the old-rich target who engaged in high warmth behavior, the semantic category "the target helped a blind woman cross the street because of external benefit (he/she wanted to get a tip)" suggests that the participant made an external attribution rather than an internal attribution for the targets' helpful behavior; that is, the participant viewed the target as showing low warmth, and the attribution would be given a low score (1) on the attribution positivity scale. By contrast, the semantic category "the target helped a blind woman cross the street because of personal internal attributes (he/she is a very warm and friendly person)" suggests that participant attributed the target's helpful behavior to internal rather than external causes; that is, the participant viewed the target as showing high warmth, and the attribution would receive a high score (5) on the attribution positivity scale. These ratings were used as guides for specific scoring of each participant's attributions.

In the third step, two other postgraduates who majored in psychology and were blind to the hypotheses of the study converted all of the participants' handwritten attributions into

numeric values (1–5) according to the coding manual made in step two. The sum of the two raters' scores was the final score of the attribution positivity, and we calculated Kendall coefficients to demonstrate the inter-rater reliability. For the low warmth behavior, the Kendall coefficients were 0.95 (rich), 0.91 (old), and 0.95 (old-rich). For the high warmth behavior, the Kendall coefficients were 0.86 (rich), 0.81 (old), and 0.80 (old-rich). For the low competence behavior, the Kendall coefficients were 0.81 (poor), 0.84 (young), and 0.79 (young-poor). For the high competence behavior, the Kendall coefficients were 0.86 (poor), 0.86 (young), and 0.74 (young-poor). This method of rating attributions has been demonstrated as valid (Song et al., in press), and the attribution positivity was used as an indirect index of stereotype evaluation.

Results

Correlation Analysis

The correlation analysis results can be seen in **Tables 4**, **5**. The results indicated that the warmth evaluation of the old-rich target was significantly correlated with the warmth evaluation of the rich and old targets in both high and low warmth scenarios. The young-poor target competence evaluation was also significantly correlated with the young and poor target competence evaluations in both high and low competence scenarios.

In the low warmth scenario, attributions for the old-rich target's behavior were positively correlated with attributions for the old target's behavior. In the high warmth scenario, however, attributions for the old-rich target's behavior were positively correlated with attributions for the rich target's behavior. Moreover, attributions for the young-poor target's behavior were significantly correlated with those for the young and poor targets' behavior in the low competence scenario, but only significantly correlated with those for the poor target's behavior in the high competence scenario.

The Moderating Role of Scenario in the Relations among Warmth Evaluations of the Old, Rich, and Old-Rich Targets

Hierarchical linear models were conducted to explore the moderating role of the scenario in the relations among the warmth evaluations of the old, rich, and old-rich targets. The first block included the warmth evaluation of the old target, the warmth evaluation of the rich target, and the scenario. Scenario was a dummy variable, with the low warmth or low competence scenario assigned 0, and the high warmth or high competence scenario assigned 1. The second block included two interaction terms, which were computed as the product of scenario and the mean-centered measure of the warmth evaluation of the old or rich target. The third block included the product term of the three independent variables. As shown in **Table 6**, in the second model, the product term of scenario and old target warmth evaluation was significant ($\beta = -0.17$, $p < 0.05$). To further examine this two-way interaction, follow-up regressions were conducted for both the high and low warmth scenarios.

As can be seen in **Table 7**, in the low warmth scenario, the results of the linear regression indicated that, after controlling for the demographic variables, the *old* warmth evaluation and the *rich* warmth evaluation accounted for significant variance in the old-rich warmth evaluation ($\beta = 0.49$, $p < 0.001$; $\beta = 0.26$, $p < 0.01$), and the relative weight of the *old* category ($RW = 0.28$) was greater than that of the *rich* category ($RW = 0.14$). In regard to the high warmth scenario, the warmth evaluations of the *old* and the *rich* targets both significantly predicted the old-rich warmth evaluation ($\beta = 0.22$, $p < 0.05$; $\beta = 0.44$, $p < 0.001$), and the relative weight of the *rich* category ($RW = 0.20$)

TABLE 4 | Correlations among warmth evaluations and attributions about the behavior of the old, rich, and old-rich targets (*N* = 95).

	Warmth evaluation						Attribution					
	Low warmth behavior			High warmth behavior			Low warmth behavior			High warmth behavior		
	1. Old	2. Rich	3. Old rich	4. Old	5. Rich	6. Old rich	7. Old	8. Rich	9. Old rich	10. Old	11. Rich	12. Old rich
1	–											
2	0.46**	–										
3	0.61**	0.48**	–									
4	0.04	−0.002	0.06	–								
5	0.04	0.08	0.12	0.44**	–							
6	−0.02	0.03	0.20	0.38**	0.51**	–						
7	0.42**	0.18	0.18	−0.13	−0.02	−0.07	–					
8	0.22*	0.48**	0.23*	−0.01	0.22*	−0.04	0.23*	–				
9	0.15	0.28**	0.54**	−0.05	−0.01	0.04	0.33**	0.10	–			
10	−0.14	−0.09	−0.11	0.19	0.01	0.09	−0.02	−0.06	−0.07	–		
11	−0.04	−0.01	0.07	0.13	0.40**	0.20*	−0.15	0.07	−0.11	0.15	–	
12	0.01	0.06	0.06	−0.14	0.01	0.18	−0.10	−0.12	−0.05	0.12	0.28**	–
M	8.54	6.61	7.36	13.01	12.12	12.56	8.37	5.06	6.57	9.37	8.76	9.25
SD	2.25	2.22	2.89	1.67	1.98	2.23	1.99	2.89	3.00	0.97	2.34	1.64

*$p < 0.05$, **$p < 0.01$.

was greater than that of the *old* category ($RW = 0.09$). H3a and H3b were supported. The category with stereotype-inconsistent behavior was the dominant category.

The Moderating Role of Scenario in the Relations among Attribution Positivity Ratings of the Old, Rich, and Old-Rich Targets

In order to explore the moderator role of the scenario in the relations between the attribution positivity ratings of the rich and old targets and the attribution positivity ratings of the old-rich targets, hierarchical linear models were conducted. As can be seen **Table 6**, the results showed that the product term of the scenario and the attribution for the old target's behavior was marginally significant ($\beta = -0.15$, $p = 0.08$). We conducted two follow-up regressions for the high and low warmth behaviors to further explore this two-way interaction.

As shown in **Table 7**, in the low warmth scenario, the regression analysis indicated that, after controlling for the demographic variables, attributions for the old target's behavior significantly positively predicted attributions for the old-rich target's behavior ($\beta = 0.33$, $p < 0.01$); however, attributions for the rich target's behavior did not predict attributions for the old-rich target's behavior ($\beta = -0.002$, $p > 0.05$). Thus, the relative weight of the *old* category ($RW = 0.10$) was greater than that of the *rich* category ($RW = 0.01$). In the high warmth scenario, attributions for the rich target's behavior marginally positively predicted attributions for the old-rich target's behavior ($\beta = 0.18$, $p < 0.1$); however, attributions for the old target's behavior did not predict attributions for the old-rich target's behavior ($\beta = 0.09$, $p > 0.05$). Thus the relative weight of the *rich* category ($RW = 0.07$) was greater than that of the *old* category ($RW = 0.01$). H3a and H3b were also supported.

The Moderating Role of Scenario in the Relations among Competence Evaluations of the Young, Poor, and Young Poor-Targets

In order to analyze how the scenario moderated the relations between the competence evaluations of the young and poor targets and the competence evaluation of the young-poor target, hierarchical linear models were used. As can be seen in **Table 6**, the results of this third model indicated that the product term of the three variables was significant ($\beta = -0.19$, $p < 0.05$). Follow-up linear regressions were conducted for both the high and low competence scenarios to further examine this three-way interaction.

As can be seen in **Table 7**, for the low competence scenario, the results of the linear regression indicated that, after controlling for the demographic variables, the competence evaluation of the poor target was a significant predictor of the young-poor competence evaluation ($\beta = 0.77$, $p < 0.001$), but the competence evaluation of the young target was not a significant predictor ($\beta = 0.04$, $p > 0.05$). Thus, the *poor* category had a greater relative weight ($RW = 0.53$). For the high competence scenario, the regression analysis showed that the competence evaluations of the young and the poor targets were both significantly positively associated with the young-poor competence evaluation ($\beta = 0.30$, $p < 0.01$; $\beta = 0.52$, $p < 0.001$). The relative weight of the *poor* category ($RW = 0.36$) was greater than that of the *young* category ($RW = 0.23$). H3c was not supported, but H3d was supported in the competence evaluations of the young-poor target. However, the competence evaluation of old-rich target was not tested, and so the dominance of the category that contradicted the stereotype expectation in the competence evaluation of the old-rich target could not be confirmed.

The Moderating Role of Scenario in the Relations among Attribution Positivity Ratings of the Young, Poor, and Young-Poor Targets

Hierarchical linear models were conducted to explore the moderator effect of scenario in the relation between the attribution positivity ratings of the young and poor targets and the attribution positivity ratings of the young-poor target. As shown in **Table 6**, the product term was not significant in either Model 2 or Model 3. The Model 1 indicated that attributions for the *poor* target's behavior and attributions for *young* target's behavior could predict attribution for the young-poor target' behavior ($\beta = 0.41$, $p < 0.001$; $\beta = 0.25$, $p < 0.01$). The relative weight analysis indicated that the *poor* category ($RW = 0.38$) was relatively more important than *young* category ($RW = 0.25$). H3c was not supported, but H3d was supported in the attributions made for the young-poor target's behavior. However, attributions for the high or low competence of the old-rich target were not tested, and so the dominance of the category that contradicted the stereotype expectation in the indirect competence evaluation of the old-rich target could not be confirmed.

Discussion of Study 2

There was some evidence that the scenario (high vs. low warmth behavior) moderated the functional significance of the young and poor categories in the explicit competence evaluation. In the high competence scenario, *poor* was the dominant category. In the low competence scenario, the relative weight of the *poor* category was less than in the high competence scenario, but was still dominant. However, the moderator effect was not supported by the indirect attribution measurement. This may be because the attribution measure, as an indirect indicator of attitude, is not sensitive enough to detect the moderator effect. With regard to the functional significance of the *old* and *rich* categories, the scenario specificity of the results was verified both in the direct warmth or competence evaluations and in the indirect attribution positivity scores. The two methods obtained relatively consistent results: The *old* category was the dominant category in the perception of the old-rich target's low warmth behavior, whereas the *rich* category was the dominant category in the perception of the old-rich target's high warmth behavior. There were also some additional findings in the regression model. Specifically, for the warmth evaluation of the old-rich target, the warmth evaluation of the old and rich were both significant predictors. In contrast, for the attribution measure, only the attribution for the old target's behavior could predict the attribution for the old-rich target's behavior in the low warmth scenario, and only the attribution for the rich target's behavior was a significant predictor in the high warmth scenario. We suspect that the perceiver would be likely to evaluate the target based on the

TABLE 5 | Correlations among competence evaluations and attributions about the behavior of the young, poor, and young-poor targets (*N* = 61).

	Competence evaluation						Attribution					
	Low competence			High competence			Low competence			High competence		
	1. Young	2. Poor	3. Young poor	4. Young	5. Poor	6. Young poor	7. Young	8. Poor	9. Young poor	10. Young	11. Poor	12. Young poor
1	–											
2	0.36**	–										
3	0.35**	077**	–									
4	0.03	0.30*	0.27*	–								
5	−0.15	0.22	0.08	0.50**	–							
6	0.03	0.08	0.08	0.61**	0.71**	–						
7	0.07	−0.18	−0.24	0.01	0.11	0.12	–					
8	−0.07	0.20	0.18	0.10	0.08	−0.01	0.01	–				
9	−0.17	0.20	0.41**	0.05	0.10	0.07	0.29*	0.39*	–			
10	0.22	−0.02	0.10	−0.12	0.07	−0.02	−0.24	−0.06	−0.08	–		
11	0.03	0.15	0.23	0.21	0.19	0.29*	−0.07	0.19	0.19	0.09	–	
12	0.16	0.16	0.27*	−0.08	−0.13	−0.01	−0.10	−0.21	0.09	0.24	0.34*	–
M	9.38	9.00	8.34	12.49	12.30	12.77	6.72	5.57	5.95	9.56	8.62	9.13
SD	1.59	1.83	2.04	1.51	1.49	1.64	1.91	1.58	1.74	0.96	0.90	0.92

*$p < 0.05$, **$p < 0.01$.

two given categories, and the weight of the two categories would determine perception of the crossed-category in the explicit evaluation. However, in the implicit evaluation, the perceiver may only perceive the target based on one dominant category, and the cognition process in the implicit evaluation would be likely to take the shortcut because it is simpler and more concise.

There were some limitations in Study 2. First, only one low warmth scenario, one high warmth scenario, one low competence scenario, and one high competence scenario were used. Although a rigorous process was performed for choosing the appropriate scenarios, the suitability and feasibility of these scenarios still need to be established. Moreover, the rigorous procedure of selecting scenario settings reduces the ecological validity and generalization of the conclusions. Second, it is still necessary to explore whether the causal attribution (external/internal) could be mixed up with trait ascription. For example, an external attribution for high warmth behavior might reflect less trait ascription (in this case "warm") than an internal attribution for high warmth behavior. The analysis of the attribution positivity requires careful consideration in specific contexts.

GENERAL DISCUSSION

The purpose of the current study was to test the functional significance of conflicting stereotypes (i.e., old-rich and young-poor), and to identify the dominant category and the weaker category in these cross-categorizations. The pilot study used a categorization task to verify that these were conflicting categories by identifying perceptions of these categories in relation to warmth and competence. In Study 1, the participants were asked to use rating scales to evaluate the competence and warmth of targets belonging to simple and crossed categories. The results in both the pilot study and Study 1 showed that the *old* category was the dominant category in the warmth evaluation of the old-rich target, and the *poor* category was the dominant category in the competence evaluation of the young-poor target. This shows that the stereotype related to the simple-category was positively associated with the stereotype related to the crossed-category, and the category with the stronger stereotype was the dominant category in the perception of the crossed-category. Study 2 further tested the functional significance of these categories in specific scenarios, and the results varied depending on the situation-dependent behavior of the target. An old-rich target who behaves warmly is judged more in line with one's evaluations regarding the rich, whereas an old-rich target that behaves unwarmly is judged more in line with one's evaluations regarding the old. However, in the competence evaluation of the young-poor target, *poor* was the dominant category in both high and low competence scenarios. Thus, the hypothesis that the category that contradicts the stereotype expectation is potentially more salient and drives judgments is partly supported.

The *rich* category was the dominant category for an old-rich target that behaves warmly, whereas the *old* category was the dominant category for an old-rich target that behaves unwarmly. The inconsistent conclusions demonstrate the scenario-specificity in the functional significance analysis of old-rich groups (Casper et al., 2011, 2015). As was evident in our findings, it is the stereotype related to a certain category combined with behavior of the target that affects the functional significance (Crisp and Hewstone, 2006), and violations of stereotypic expectancy (the high warmth behavior of the rich target, and the low warmth behavior of the old target) attract more attention (Bettencourt et al., 1997; Dickter and Gyurovski, 2012). It may be because the category that violates the stereotype expectancy would be salient and relatively more accessible, and thus, would be selected as the dominant category. But further research is still needed to determine exactly how this process unfolds.

TABLE 6 | Hierarchical Linear Models of the moderating role of scenario in the relation between the simple-category stereotype evaluations and crossed-category stereotype evaluation.

		Warmth or competence evaluation		Attribution	
		β	*t*	*β*	*t*
Model 1	Old	0.40	5.76***	0.24	3.66***
	Rich	0.37	4.87***	0.11	1.42
	Scenario	0.11	1.57	0.35	4.65***
		$R^2 = 0.68$		$R^2 = 0.31$	
Model 2	Old	0.51	5.88**	0.28	3.88***
	Rich	0.32	3.08**	0.04	0.37
	Scenario	0.14	1.86†	0.36	4.63***
	Scenario × old	−0.17	−2.09*	−0.09	−1.22
	Scenario × rich	0.08	0.89	0.11	1.15
		$R^2 = 0.69$		$R^2 = 0.32$	
Model 3	Old	0.51	5.87***	0.28	3.89***
	Rich	0.32	3.07**	0.04	0.37
	Scenario	0.13	1.54	0.37	4.74***
	Scenario × old	−0.16	−1.59	−0.15	−1.74†
	Scenario × rich	0.08	0.65	0.06	0.61
	Scenario × rich × old	−0.01	−0.07	0.12	1.38
		$R^2 = 0.69$		$R^2 = 0.32$	
Model 1	Young	0.16	2.75**	0.24	3.27**
	Poor	0.60	10.13***	0.38	4.60***
	Scenario	0.23	3.85***	0.30	3.07**
		$R^2 = 0.83$		$R^2 = 0.66$	
Model 2	Young	0.09	1.08	0.25	3.06**
	Poor	0.67	8.99***	0.41	4.22***
	Scenario	0.24	3.96***	0.33	2.98**
	Scenario × young	0.12	1.59	−0.03	−0.32
	Scenario × poor	−0.11	−1.50	−0.05	−0.48
		$R^2 = 0.83$		$R^2 = 0.67$	
Model 3	Young	0.09	1.11	0.25	3.05**
	Poor	0.67	9.19***	0.41	4.20***
	Scenario	0.20	3.43**	0.28	1.96†
	Scenario × young	0.24	2.76**	0.03	0.19
	Scenario × poor	−0.03	−0.42	0.01	0.05
	Scenario × young × poor	−0.19	−2.50*	−0.07	−0.44
		$R^2 = 0.84$		$R^2 = 0.67$	

Scenario was a dummy variable, with the low warmth or competence scenario was 0, high warmth or competence scenario was 1. Interaction terms were computed as the product of scenario and the mean-centered measure of the warmth/competence evaluation of the simple category target(s). †*p < 0.1,*p < 0.05, **p < 0.01, ***p < 0.001.*

In the competence evaluation of the young-poor target, we found *poor* was the primary category when the behavior information was not given. Consistent with our hypotheses, the category with the stronger stereotype was the dominant category, and the low competence stereotype of the poor target was much stronger than the high competence stereotype of the young target. When considering the information about the target's behavior, we obtained consistent results with *poor* always being the dominant category. The dominance of the stereotype-inconsistent category in evaluation of cross-categorized targets was shown in the warmth-evaluation of the old-rich targets, but not found in the competence-evaluation of the young-poor targets. This might be due to the content of the dependent variable (warmth or competence) as well as the stereotype content (stereotype

TABLE 7 | Hierarchical Linear Models of evaluations of simple-category targets in relation to evaluations of crossed-category targets in specific scenarios.

Dependent variables	Independent variables	Low warmth or competence			High warmth or competence		
		β	*t*	*RW*	*β*	*t*	*RW*
Warmth evaluation of the old-rich target	Age	0.03	0.37		0.11	1.11	
	Gender	−0.06	−0.65		0.09	0.95	
	Wealth	−0.11	−1.21		0.27	2.83**	
	Hukou	0.04	0.49		−0.07	−0.74	
	Old	0.49	5.43***	0.28 (66.18%)	0.22	2.24*	0.09 (29.93%)
	Rich	0.26	2.83**	0.14 (33.82%)	0.44	4.49***	0.20 (70.07%)
		$R^2 = 0.44$			$R^2 = 0.36$		
Attributions about the old-rich target	Age	−0.10	−0.91		−0.33	−0.30	
	Gender	−0.00	−0.02		−0.13	−1.20	
	Wealth	−0.16	−1.45		−0.08	−0.74	
	Hukou	−0.08	−0.70		0.24	2.10*	
	Old	0.33	3.12**	0.10 (94.53%)	0.09	0.88	0.01 (12.92%)
	Rich	−0.00	−0.02	0.01 (5.47%)	0.18	1.71†	0.07 (87.08%)
		$R^2 = 0.15$			$R^2 = 0.14$		
Competence evaluation of the young-poor target	Age	0.31	3.56**		0.19	1.97†	
	Gender	0.03	0.32		0.07	0.70	
	Wealth	−0.06	−0.67		−0.03	−0.30	
	Hukou	0.13	1.40		−0.07	−0.71	
	Young	0.04	0.43	0.06 (10.86%)	0.30	3.11**	0.23 (38.67%)
	Poor	0.77	8.57***	0.53 (89.14%)	0.52	5.09***	0.36 (61.33%)
		$R^2 = 0.68$			$R^2 = 0.62$		

*RW = raw relative weights, and numbers in brackets refer to rescaled relative weight estimates reported as percentage of predicted variance. Hukou is a household registration system in China, and it includes two types: rural and city. †p < 0.1,*p < 0.05, **p < 0.01, ***p < 0.001.*

evaluation of the rich, old, young, and poor targets). We suspect that when there was a stronger stereotype of a certain category, any moderator effect of the additional behavior information would be lessened. The *poor* category dominated the competence ratings relatively independent of scenario. Regardless of the young-poor target's behavior, the perceiver would be likely to evaluate the target based on the *poor* category.

There is also another possibility. Most of the participants in Studies 1 and 2 saw themselves as "average" (neither poor nor rich), and thus the *poor* as well as the *rich* category constituted the out-group for most. However, most participants were young and may have seen themselves as in-group members of the *young* category. The results showing that additional behavioral information affected the evaluation of the old-rich target, but not the young-poor target, may have occurred because that stereotype-inconsistent information only dominates the evaluation of cross-categorized targets if the stereotype relates to an out-group category, but not to an in-group category. Further research is needed to verify this assumption.

It should be noted that participants may know examples of particular subtypes of persons, for example philanthropic old, rich people who are warm and caring, and young college students who came from a poor family, but have ambition and ability. Therefore, subtypes, rather than superordinate categories, may be driving participants' decisions. In addition, when additional identities like gender and race are not specified, participants may impose "male" and "in-group" identities on to the targets they are imagining (Cuddy et al., 2015). There also is the possibility that perceptions may be specific to old-rich men or old-rich women based on the previous experience of the perceiver, and the old-rich men may be evaluated differently from old-rich women in terms of stereotype-based assumptions about how targets acquired their wealth. Perceivers would make a positive competence evaluation of the old-rich men if they make the stereotypic assumption that old-rich men earn the wealth themselves, but may make a negative competence evaluation of old-rich women if they assume they acquired the wealth through a relationship with a rich partner.

Although we acknowledge that we have not conclusively pinpointed the mechanism underlying the functional significance of conflicting stereotypes, our studies do suggest some clues. We explored the co-effect of the stereotype related to a certain category and the behavior of the target. The results indicated that, in the warmth evaluation of the old-rich target, the category that showed behavior contrary to the stereotype expectation was the dominant category, but this was not found in the competence evaluation of the young-poor target. These findings

are partly consistent with previous research showing that the target whose behavior violates the stereotype expectation attracts more attention, but we have extended this research by demonstrating the effect of stereotype expectation violation on the functional significance of the conflicting old and rich categories in the warmth evaluation. In addition, we extended research on the effect of stereotype expectation violation to examine context as a moderator of this effect. Here, there is no theoretical explanation for why context in the form of target behaviors moderated the salience of some categories but not others, although we speculate that this effect may disappear in the in-group evaluation, or be reduced when there is a very strong stereotype related to warmth or competence of the category. From a more practical and applied point of view, knowledge of functional significance obtained from the current study can be utilized to help individuals to find more effective intervention strategies designed to reduce prejudice. It will be important in future research to determine whether intervention targeting the dominant category or the "other" category will be most useful for reducing stereotypes.

We also make a methodical contribution. The pilot study used a categorization task to identify the functional significance of the conflicting stereotypes, and Study 2 extended the stereotypic explanatory bias approach to study perceptions of cross-categorization groups based on participants' attributions. Sekaquaptewa et al. (2003) posited that subtracting the number of explanations (internal or external) for stereotype-consistent events from the number of explanations provided for stereotype-inconsistent events provides an indirect measure of the stereotype. In contrast, in the current study we coded the attribution based on the attribution positivity, which is a more sensitive index compared with the type of attribution (internal or external attribution).

With regard to other limitations and potential extensions of the current work there are several issues worthy of note. First, members of an "out-group" based on one category may be evaluated more positively if they are also members of an "in-group" based on another category. The sample in the current study was made up of young, educated, mostly female participants, most of whom probably identify themselves as members of the *young* category. As this may influence the results, further research should include older participants, so as to take into consideration the role of the intergroup identity. Future research on this issue is important, as in-group identification is one important mechanism for reducing prejudice and defamation against a cross-categorized group (Crisp et al., 2003; Ray et al., 2010). Second, further study should also focus on sub-categories like middle-class (rather than rich and poor) and middle-aged (as opposed to old and young) targets. Moreover, other categories such as gender and non-dichotomous categories such as race (African American, Asian, White, etc.) need more attention. Third, the current research only focused on conflicting cross-categorization, but an analysis of the functional significance of consistent cross-categorization might also prove valuable, and future research on this topic is needed.

ETHICS STATEMENT

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee, the American Psychological Association (APA) standards and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

AUTHOR CONTRIBUTIONS

JS conceived and designed the experiments, analyzed and interpreted the data, wrote the report. BZ had role in the study design.

ACKNOWLEDGMENTS

The authors sincerely thank the reviewers and editor for many productive suggestions during the interactive review. This research was supported by grants from a general program grant 31571147 from National Natural Science Foundation of China, grants from Young Scholar Grant 31400903 from National Natural Science Foundation of China, and Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE grant CCNU15Z02001.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01624

REFERENCES

Abele, A. E., and Wojciszke, B. (2014). Communal and agentic content in social cognition: a dual perspective model. *Adv. Exp. Soc. Psychol.* 50, 195–255. doi: 10.1016/B978-0-12-800284-1.00004-7

Bettencourt, B. A., Dill, K. E., Greathouse, S. A., Charlton, K., and Mulholland, A. (1997). Evaluations of ingroup and outgroup members: the role of category-based expectancy violation. *J. Exp. Soc. Psychol.* 33, 244–275. doi: 10.1006/jesp.1996.1323

Biernat, M., and Vescio, T. K. (2002). She swings, she hits, she's great, she's benched: implications of gender-based shifting standards for judgment and behavior. *Pers. Soc. Psychol. Bull.* 28, 66–77. doi: 10.1177/0146167202281006

Bodenhausen, G. V. (2010). Diversity in the person, diversity in the group: challenges of identity complexity for social perception and social interaction. *Eur. J. Soc. Psychol.* 40, 1–16. doi: 10.1002/ejsp.647

Burgers, C., and Beukeboom, C. J. (2014). Stereotype transmission and maintenance through interpersonal communication the irony bias. *Commun. Res.* 43, 414–441. doi: 10.1177/0093650214534975

Casper, C., Rothermund, K., and Wentura, D. (2011). The activation of specific facets of age stereotypes depends on individuating

information. *Soc. Cogn.* 29, 393–414. doi: 10.1521/soco.2011.29.4.393

Casper, C., Rothermund, K., and Wentura, D. (2015). Automatic stereotype activation is context dependent. *Soc. Psychol.* 41, 131–136. doi: 10.1027/1864-9335/a000019

Chasteen, A., Kang, S., and Remedios, J. (2012). "Aging and stereotype threat: development, process, and interventions," in *Stereotype Threat: Theory, Process, and Application*, Vol. 2, eds M. Inzlicht and T. Schmader (New York, NY: Oxford University Press), 202–216.

Cloutier, J., Freeman, J. B., and Ambady, N. (2014). Investigating the early stages of person perception: the asymmetry of social categorization by sex vs. age. *PLoS ONE* 9:e84677. doi: 10.1371/journal.pone.0084677

Corcoran, K., Hundhammer, T., and Mussweiler, T. (2009). A tool for thought! When comparative thinking reduces stereotyping effects. *J. Exp. Soc. Psychol.* 45, 1008–1011. doi: 10.1016/j.jesp.2009.04.015

Crisp, R. J., and Hewstone, M. (2006). *Multiple Social Categorization*: *Context, Process, and Social Consequences*. Hove; New York, NY: Psychology Press.

Crisp, R. J., Hewstone, M., Richards, Z., and Paolini, S. (2003). Inclusiveness and crossed categorization: effects on co-joined category evaluations of in-group and out-group primes. *Br. J. Soc. Psychol.* 42, 25–38. doi: 10.1348/014466603763276108

Cuddy, A. J., Fiske, S. T., Kwan, V. S., Glick, P., Demoulin, S., Leyens, J. P., et al. (2009). Stereotype content model across cultures: towards universal similarities and some differences. *Br. J. Soc. Psychol.* 48, 1–33. doi: 10.1348/014466608X314935

Cuddy, A. J., Wolf, E. B., Glick, P., Crotty, S., Chong, J., and Norton, M. I. (2015). Men as cultural ideals: cultural values moderate gender stereotype content. *J. Pers. Soc. Psychol.* 109, 622–635. doi: 10.1037/pspi0000027

Dickter, C., and Gyurovski, I. (2012). The effects of expectancy violations on early attention to race in an impression-formation paradigm. *Soc. Neurosci.* 7, 240–251. doi: 10.1080/17470919.2011.609906

Dommelen, A., Schmid, K., Hewstone, M., Gonsalkorale, K., and Brewer, M. (2015). Construing multiple ingroups: assessing social identity inclusiveness and structure in ethnic and religious minority group members. *Eur. J. Soc. Psychol.* 45, 386–399. doi: 10.1002/ejsp.2095

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., and Howard, A. (1997). On the nature of prejudice: automatic and controlled processes. *J. Exp. Soc. Psychol.* 33, 510–540. doi: 10.1006/jesp.1997.1331

Eich, T. S., Murayama, K., Castel, A. D., and Knowlton, B. J. (2014). The dynamic effects of age-related stereotype threat on explicit and implicit memory performance in older adults. *Soc. Cogn.* 32, 559–570. doi: 10.1521/soco.2014.32.6.559

Fazio, R. H., Jackson, J. R., Dunton, B. C., and Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *J. Pers. Soc. Psychol.* 69, 1013–1027. doi: 10.1037/0022-3514.69.6.1013

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* 82, 878–896. doi: 10.1037/0022-3514.82.6.878

Freeman, J. B., and Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychol. Rev.* 118, 247–249. doi: 10.1037/a0022327

Garcia-Marques, T., Mackie, D. M., Maitner, A. T., and Claypool, H. M. (2016). Moderation of the familiarity-stereotyping effect: the role of stereotype fit. *Soc. Cogn.* 34, 81–96. doi: 10.1521/soco.2016.34.2.81

Hess, T. M., Hinson, J. T., and Hodges, E. A. (2009). Moderators of and mechanisms underlying stereotype threat effects on older adults' memory performance. *Exp. Aging Res.* 35, 153–177. doi: 10.1080/03610730802716413

Jerónimo, R., Volpert, H. I., and Bartholow, B. D. (2016). Event-related potentials reveal early attention bias for negative, unexpected behavior. *Soc. Neurosci.* doi: 10.1080/17470919.2016.1144646. [Epub ahead of print].

Johnson, K. L., Freeman, J. B., and Pauker, K. (2012). Race is gendered: how covarying phenotypes and stereotypes bias sex categorization. *J. Pers. Soc. Psychol.* 102, 116–131. doi: 10.1037/a0025335

Judd, C. M., James-Hawkins, L., Yzerbyt, V., and Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *J. Pers. Soc. Psychol.* 89, 899–913. doi: 10.1037/0022-3514.89.6.899

Kang, S. K., and Bodenhausen, G. V. (2015). Multiple identities in social perception and interaction: challenges and opportunities. *Annu. Rev. Psychol.* 66, 547–574. doi: 10.1146/annurev-psych-010814-015025

Kang, S. K., Chasteen, A. L., Cadieux, J., Cary, L. A., and Syeda, M. (2014). Comparing young and older adults' perceptions of conflicting stereotypes and multiply-categorizable individuals. *Psychol. Aging* 29, 469–481. doi: 10.1037/a0037551

Kite, M. E., Stockdale, G. D., Whitley, B. E., and Johnson, B. T. (2005). Attitudes toward younger and older adults: an updated meta-analytic review. *J. Soc. Issues* 61, 241–266. doi: 10.1111/j.1540-4560.2005.00404.x

Klauer, K. C., Ehrenberg, K., and Wegener, I. (2003). Crossed categorization and stereotyping: structural analyses, effect patterns, and dissociative effects of context relevance. *J. Exp. Soc. Psychol.* 39, 332–354. doi: 10.1016/S0022-1031(03)00017-9

Klauer, K. C., Hölzenbein, F., Calanchini, J., and Sherman, J. W. (2014). How malleable is categorization by race? Evidence for competitive category use in social categorization. *J. Pers. Soc. Psychol.* 107, 21–40. doi: 10.1037/a0036609

Kunda, Z., Sinclair, L., and Griffin, D. (1997). Equal ratings but separate meanings: stereotypes and the construal of traits. *J. Pers. Soc. Psychol.* 72, 720–734. doi: 10.1037/0022-3514.72.4.720

LeBreton, J. M., and Tonidandel, S. (2008). Multivariate relative importance: extending relative weight analysis to multivariate criterion spaces. *J. Appl. Psychol.* 93, 329–345. doi: 10.1037/0021-9010.93.2.329

Ma, D. S., and Correll, J. (2011). Target prototypicality moderates racial bias in the decision to shoot. *J. Exp. Soc. Psychol.* 47, 391–396. doi: 10.1016/j.jesp.2010.11.002

Neuberg, S. L., and Sng, O. (2013). A life history theory of social perception: stereotyping at the intersections of age, sex, ecology (and race). *Soc. Cogn.* 31, 696–711. doi: 10.1521/soco.2013.31.6.696

Pendry, L. F., and Macrae, C. N. (1996). What the disinterested perceiver overlooks: goal-directed social categorization. *Pers. Soc. Psychol. Bull.* 22, 249–256. doi: 10.1177/0146167296223003

Penner, A. M., and Saperstein, A. (2013). Engendering racial perceptions an intersectional analysis of how social status shapes race. *Gender Soc.* 27, 319–344. doi: 10.1177/0891243213480262

Piff, P. K., Kraus, M. W., Côté, S., Cheng, B. H., and Keltner, D. (2010). Having less, giving more: the influence of social class on prosocial behavior. *J. Pers. Soc. Psychol.* 99, 771–784. doi: 10.1037/a0020092

Ray, D. G., Way, N., and Hamilton, D. L. (2010). Crossed-categorization, evaluation, and face recognition. *J. Exp. Soc. Psychol.* 46, 449–452. doi: 10.1016/j.jesp.2009.12.001

Schug, J., Alt, N. P., and Klauer, K. C. (2015). Gendered race prototypes: evidence for the non-prototypicality of Asian men and Black women. *J. Exp. Soc. Psychol.* 56, 121–125. doi: 10.1016/j.jesp.2014.09.012

Sekaquaptewa, D., and Espinoza, P. (2004). Biased processing of stereotype-incongruency is greater for low than high status groups. *J. Exp. Soc. Psychol.* 40, 128–135. doi: 10.1016/S0022-1031(03)00093-3

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., and von Hippel, W. (2003). Stereotypic explanatory bias: implicit stereotyping as a predictor of discrimination. *J. Exp. Soc. Psychol.* 39, 75–82. doi: 10.1016/S0022-1031(02)00512-7

Sesko, A. K., and Biernat, M. (2010). Prototypes of race and gender: the invisibility of Black women. *J. Exp. Soc. Psychol.* 46, 356–360. doi: 10.1016/j.jesp.2009.10.016

Singh, R., Yeoh, B. S., Lim, D. I., and Lim, K. K. (1997). Cross-categorization effects in intergroup discrimination: adding versus averaging. *Br. J. Soc. Psychol.* 36, 121–138. doi: 10.1111/j.2044-8309.1997.tb01123.x

Smith, E. R., Fazio, R. H., and Cejka, M. A. (1996). Accessible attitudes influence categorization of multiply categorizable objects. *J. Pers. Soc. Psychol.* 71, 888–898. doi: 10.1037/0022-3514.71.5.888

Smith, E. R., Stewart, T. L., and Buttram, R. T. (1992). Inferring a trait from a behavior has long-term, highly specific effects. *J. Pers. Soc. Psychol.* 62, 753–759. doi: 10.1037/0022-3514.62.5.753

Song, J., Zuo, B., Wen, F., Tan, X., and Zhao, M. (in press). The effect of crossed categorization on stereotype-wealth × age group. *Psychol. Explor.*

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. S. (1987). *Rediscovering the Social Group*: *A Self-Categorization Theory*. Oxford: Basil Blackwell.

Urada, D., Stenstrom, D. M., and Miller, N. (2007). Crossed categorization beyond the two-group model. *J. Pers. Soc. Psychol.* 92, 649–664. doi: 10.1037/0022-3514.92.4.649

Van Rijswijk, W., and Ellemers, N. (2002). Context effects on the application of stereotype content to multiple categorizable targets. *Pers. Soc. Psychol. Bull.* 28, 90–101. doi: 10.1177/0146167202281008

Conflict of Interest Statement: 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 reviewer PS and the handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2016 Song and Zuo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 24 May 2017
doi: 10.3389/fpsyg.2017.00857

How Do Discrepancies between Victimization and Rejection Expectations in Gay and Bisexual Men Relate to Mental Health Problems?

Frank A. Sattler and Hanna Christiansen*

Department of Clinical Child and Adolescent Psychology, Philipps University of Marburg, Marburg, Germany

Introduction: Victimization and rejection expectations predict mental health problems in gay and bisexual men. Furthermore, it was shown that victimization predicts rejection expectations. Nevertheless, the levels of these two variables do not necessarily correspond as indicated by low inter-correlations, resulting in the question "How do discrepancies in the two variables relate to mental health problems?" This study tests if non-corresponding levels of victimization and rejection expectations in gay and bisexual men relate to mental health problems differently than corresponding levels of victimization and rejection expectations. It furthermore tests for linear and curvilinear relationships between victimization, rejection expectations, and mental health problems.

Methods: Data from $N = 1423$ gay and bisexual men were obtained online. Victimization and rejection expectations were tested for discrepant values (differing 0.5 SD or more) and those that were in agreement (differing less than 0.5): 33.7% of participants were in agreement, 33.0% reported higher rejection expectations than victimization, and 33.3% v.v. Then, a polynomial regression and a surface analysis were conducted.

Results: Discrepant values in victimization and rejection expectations or the direction of the discrepancy did not relevantly predict mental health problems. Findings indicate that victimization and rejection expectations predict mental health problems linearly as well as convexly (upward curving) in gay and bisexual men.

Discussion: This study replicates findings that gay and bisexual men with more experiences of victimization and rejection expectations demonstrated more mental health problems. Furthermore, this study is the first one to find a convex relationship between these predictors and mental health problems, implicating that disproportionally high mental health problems exist in those gay and bisexual men with high levels of victimization and rejection expectations. On the other hand, discrepancies between these two variables do not predict mental health problems. Future studies are needed to test for replication of our findings.

Keywords: rejection expectations, victimization, expectation violations, mental health, gay and bisexual men

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Charlotte Tate, San Francisco State University, United States
Franziska Labrenz, Essen University Hospital, Germany

****Correspondence:***
Frank A. Sattler
frank.sattler@uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *01 October 2016*
Accepted: *09 May 2017*
Published: *24 May 2017*

Citation:
Sattler FA and Christiansen H (2017) How Do Discrepancies between Victimization and Rejection Expectations in Gay and Bisexual Men Relate to Mental Health Problems? Front. Psychol. 8:857. doi: 10.3389/fpsyg.2017.00857

INTRODUCTION

According to minority stress theory, a number of minority stressors lead to mental health problems in gay and bisexual men, resulting in mental health disparities between gay and bisexual men in comparison to heterosexual men (Meyer, 2003; King et al., 2008). Minority stressors faced by gay and bisexual men include gay-related victimization, discrimination, rejection expectations (chronic expectations of gay-related rejection), internalized homonegativity (or internalized homophobia), boyhood gender non-conformity, and masculine standards (Pachankis, 2015). It is proposed that these stressors lead to a higher number of mental health problems as other non-minority specific stressors (e.g., work stress or marital stress) would do, too. Among the minority stressors with the broadest empirical evidence are gay-related victimization (i.e., victimization of gay and bisexual men due to their sexual orientation) and rejection expectations (i.e., expectation of being a target of victimization in the future). Numerous cross-sectional studies and some longitudinal ones have found that these minority stressors linearly predict gay and bisexual men's mental health problems (Feinstein et al., 2012; Burton et al., 2013; Pachankis et al., 2014a; Eldahan et al., 2016; Sattler et al., 2016). Up to date no studies exist that tested for a curvilinear (squared) relationship between these variables. The knowledge is thus very limited on how both variables might interact with one another.

Furthermore, it was proposed that minority stressors are not independent from each other but that gay-related victimization (from now on abbreviated as victimization) predicts expectations such as rejection expectations (Fredriksen-Goldsen et al., 2014). Indeed, cross-sectionally it was demonstrated that victimization predicted rejection expectations in lesbians and gay men (Feinstein et al., 2012). Nevertheless, the level of rejection expectations does not necessarily correspond to the level of victimization in each gay or bisexual man as shown by studies reporting low associations ($r = 0.20$ to 0.29) between the two variables (Pachankis et al., 2014b; Sattler et al., 2016). There are two possible scenarios: (1) an individual may expect to be rejected although they have been victimized in the past to a non-correspondingly low degree or, (2) an individual may expect very little rejection despite having been victimized in the past to a non-correspondingly high degree. In both cases, an expectation violation is prevalent; or in other words discrepancies exist between victimization and rejection expectations.

The primary goal of the study is to empirically investigate these expectation violations. As implied by earlier studies, we therefore hypothesize that we will find a linear relationship between victimization, rejection expectations, and mental health (hypothesis 1). Furthermore, we were interested in whether the relation between victimization, rejection expectations, and mental health problems is best described as merely linear or if an interaction exists. We therefore wanted to test whether differing levels of victimization and rejection expectations will predict differing levels of mental health problems, in addition to the predictions depicted in hypothesis 1, and whether victimization and rejection expectations predict mental health problems curvilinearly (squaredly).

MATERIALS AND METHODS

Data Collection

The survey was conducted online in a number of German web sites for gay and bisexual men as well as mailing lists for students and employees of the Philipps University of Marburg (PUM). This study was carried out in accordance with the recommendations of the Ethics Committee of the Psychological Faculty of the Philipps University of Marburg (PUM) with online informed consent from all subjects. All subjects gave online informed consent in accordance with the Declaration of Helsinki. The study was approved by the Ethics Committee of the Psychological Faculty of the PUM.

Participants

In total, $N = 1737$ gay and bisexual men participated in the survey in 2014. Participants who indicated that they were younger than 18 years ($n = 3$), older than 80 years ($n = 18$), or who did not complete the questionnaire ($n = 293$) were excluded from analyses. The final sample thus consisted of $N = 1423$ gay and bisexual men. Of these men, $n = 1308$ (91.9%) defined as gay and $n = 115$ (8.1%) defined as bisexual. Furthermore, $n = 146$ (10.3%) were immigrants or had at least one immigrant parent. The relationship status was as follows: $n = 688$ (48.3%) gay and bisexual men were in a relationship with a man; $n = 158$ (11.1% of the total sample) of them were in a civil union. Furthermore, $n = 50$ (3.5%) were in a relationship with a woman; $n = 32$ (2.2% of the total sample) of them were married. Finally, $n = 691$ (48.6%) were single. The education levels were as follows: $n = 3$ (0.2%) no school degree, $n = 57$ (4.0%) junior high school degree, $n = 193$ (13.6%) middle high school degree, $n = 420$ (29.5%) senior high school degree, $n = 624$ (43.9%) university degree, and $n = 126$ (8.9%) doctoral degree.

Measures

Victimization

It was assessed with five items of the victimization scale by Herek and Berrill (1992). The items asked for victimization since the age of 16 years. While the original scale used a three-point response format (from 1 = *never* to 3 = *two or more*), we used an amplified four-point response format (from 1 = *never* to 4 = *three times or more*). Cronbach's alpha of the scale was 0.76 in the present study.

Rejection Expectations

It was assessed with three items of the Gay-Related Rejection Sensitivity Scale (Pachankis et al., 2008). The participants read three short texts on potentially homonegative situations and reported whether they would feel discriminated upon in these situations due to their sexual orientation. A five-point response format was used (from 1 = *strongly disagree* to 5 = *strongly agree*). Cronbach's alpha of the scale was 0.65 in the present study. Due to the Cronbach's alpha that was between the thresholds of questionable (0.60) and sufficient (0.70), a principal component factor analysis ($\kappa = 4$; number of iterations = 1000) was applied to test the factorial validity of the rejection expectations scale. Only one component with an eigenvalue > 1 was extracted,

thereby explaining 58.3% of the variance. All items loaded on the component between $\lambda = 0.71$ and 0.81.

Mental Health

The problems were assessed with 27 items of the Brief Symptom Inventory (Franke, 2000). The items assessed symptoms of somatization, obsessive-compulsive disorder, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism. A five-point response format was used (from 1 = *not at all* to 5 = *extremely*). The scale's Cronbach's alpha was 0.95 in the present study.

Data Analysis

For data analysis, zero-order Pearson's correlations between the main constructs were computed. Scores for rejection expectations and victimization were z-standardized. Then, a polynomial regression with response surface analysis was conducted using the approach described by Shanock et al. (2010) that includes the following steps: first, descriptive information was provided about the occurrence of discrepancies within the variables victimization and rejection expectations. Thereby, any participant with the two scores differing half a standard deviation or more were considered to have discrepant values (Shanock et al., 2010), while the rest was considered to have agreeing values for the two constructs. Second, a polynomial regression was conducted in IBM Statistics SPSS 22 and the surface values were conducted afterward. Thereby, the predictors were centered around the midpoint of their respective scales (Shanock et al., 2010). Then, the following variables were computed: the square of the centered variable victimization, the square of the centered variable rejection expectations, and the cross-product of both centered variables. Afterward, a polynomial regression was conducted using the centered predictor variables, the squared variables, and the cross-product variable as predictors. Mental health problems were used as the criterion. Third, the surface values were interpreted.

RESULTS

Descriptive Data Analysis

Victimization was positively inter-correlated with rejection expectations ($r = 0.25$, $p < 0.001$). Moreover, mental health problems were positively associated with victimization and rejection expectations ($r = 0.31$ to 0.34, $p < 0.001$). See **Table 1** for further details.

TABLE 1 | Intercorrelations of the scales.

Scale	1	2	3
(1) Victimization			
(2) Rejection expectations	0.25***		
(3) Mental health problems	0.34***	0.31***	
Mean (SD)	1.59 (0.64)	2.87 (0.98)	1.60 (0.60)
Min–Max	1–4	1–5	1–4.19

*SD, standard deviation; Min, minimum; Max, maximum. ***p < 0.001.*

Polynomial Regression with Response Surface Analysis

Step 1: Descriptive Information on Discrepancies

Data suggests the values in victimization and rejection expectations were in agreement for 33.7% of participants (meaning that they differed less than 0.5 SD), while 33.0% reported higher rejection expectations than victimization, and 33.3% reported higher victimization than rejection expectations (see **Table 2**). Since 66.3% of the predictor variables showed discrepant values, it is meaningful to use a polynomial regression for further data analysis.

Step 2: Polynomial Regression and Surface Values

The centered variable victimization ($\beta = 0.25$, $p < 0.001$), as well as the centered variable rejection expectations ($\beta = 0.18$, $p < 0.001$) significantly predicted mental health problems (see **Table 3**). Furthermore, the squared variable rejection expectations predicted mental health problems ($\beta = 0.04$, $p < 0.01$), while victimization squared did not predict mental health problems ($\beta = 0.01$, $p > 0.05$). The linear as well as the squared relationships are displayed in **Figures 1, 2**. Furthermore, the cross-product of victimization centered and rejection expectations centered significantly predicted mental health problems ($\beta = 0.05$, $p < 0.05$). However, since the predictions by rejection expectations squared and the cross-product were below $\beta < 0.10$, we interpret them as not relevant in order to not over-interpret our findings (Nathans et al., 2012). All predictors included in the polynomial regression explained 17.6% of the variance in mental health problems.

In addition, the surface values were predicted for the polynomial regression: these include the slope of the line of perfect agreement (when victimization and rejection expectations are in agreement) a_1, the curvature along the line of perfect agreement (when a squared relationship exists) a_2, the slope of the line of incongruence (when discrepancies between victimization and rejection expectations exist) a_3, as well as the curvature of the line of incongruence (when a squared relationship exists) a_4. In the current polynomial regression, a_1 ($\beta = 0.43$, $p < 0.001$), a_2 ($\beta = 0.10$, $p < 0.001$), and a_3 ($\beta = 0.07$, $p < 0.05$) proved to be significant, while a_4 ($\beta = 0.02$, $p > 0.05$) was not significant (compare **Table 3**).

Step 3: Interpretation of the Surface Values

Since a_1 was significant ($\beta = 0.43$, $p < 0.001$), there is a linear (additive) relationship between victimization, rejection expectations, and the outcome. Consequently, mental health

TABLE 2 | Agreement between victimization and rejection expectations.

Agreement groups	Percentage	Number	Mean (SD) V	Mean (SD) RE
RE more than V	33.0	470	1.27 (0.31)	3.66 (0.61)
In agreement	33.7	479	1.52 (0.48)	2.81 (0.72)
V more than RE	33.3	474	1.98 (0.79)	2.14 (0.92)

SD, standard deviation; V, victimization; RE, rejection expectations.

TABLE 3 | Discrepancy between victimization and rejection expectations as predictor of mental health problems.

Predictor	β (SE)
Victimization	0.25 (0.02)***
Rejection expectations	0.18 (0.02)***
Victimization squared	0.01 (0.03)
Rejection expectations squared	0.04 (0.01)**
Victimization × rejection expectations	0.05 (0.03)*
Surface test	
a_1	0.43***
a_2	0.10***
a_3	0.07*
a_4	0.02

*Victimization and rejection expectations are centered around the midpoint of the respective scales, β, unstandardized beta weight; SE, standard error; a_1, slope of the line of perfect agreement; a_2, curvature along the line of perfect agreement; a_3, slope of the line of incongruence; a_4, curvature of the line of incongruence. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

problems are predicted positively by agreeing levels of victimization and rejection expectations. Hypothesis 1 was therefore confirmed.

A significant a_2 ($\beta = 0.10$, $p < 0.001$) indicates that there is a non-linear slope of the line of perfect agreement. This means that the line has a convex (upward curving) surface, indicating that mental health increases to a steeper degree by increasing levels of agreeing victimization and rejection expectations. A squared relationship between these variables, was therefore not found. Both the linear as well as the squared predictions of mental health problems are depicted separately for both predictors in **Figures 1**, **2**.

Since a_3 was significant ($\beta = 0.07$, $p < 0.05$), the direction of the discrepancy is related to the outcome: mental health problems are higher when victimization exceeds rejection expectations. However, a_3 was below a level of $\beta > 0.10$ and is thus no relevant predictor of mental health problems. Furthermore, a non-significant a_4 ($\beta = 0.02$, $p > 0.05$) indicates that a stronger discrepancy does not predict a higher level of mental health problems. Therefore, no squared relationship between a discrepancy and mental health problems exists.

In summary, mental health problems were predicted linearly and furthermore convexly by agreeing levels of victimization and rejection expectations. No relevant prediction was found for discrepant values or the direction of the discrepancy.

DISCUSSION

This study is the first one to investigate if discrepancies between victimization and rejection expectations reflect on the mental health of sexual minorities.

In a sample of $N = 1423$ gay and bisexual German men, we found that agreeing levels of victimization and rejection expectations predicted mental health problems linearly as well as convexly (squaredly). Our study therefore replicates a great number of studies that found evidence of victimization and rejection expectations to predict mental health problems linearly (Frisell et al., 2010; Feinstein et al., 2012; McLaughlin et al., 2012; Burton et al., 2013; Eaton, 2014; Sattler et al., 2016). On the other hand, the findings of the squared relationship are unique: to the author's knowledge we are the first ones to demonstrate that when victimization and rejection expectations are both high, disproportionately higher levels of mental health problems are

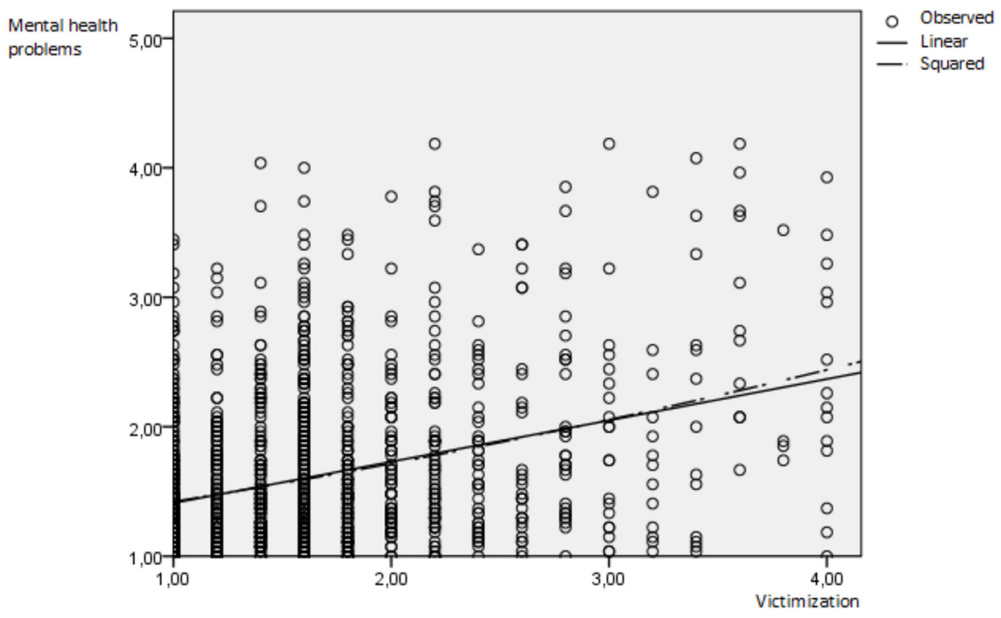


FIGURE 1 | Linear and convex prediction of mental health problems by victimization. Observed = observed value, linear = linear line of parameter estimation, squared = squared (convex) line of parameter estimation.

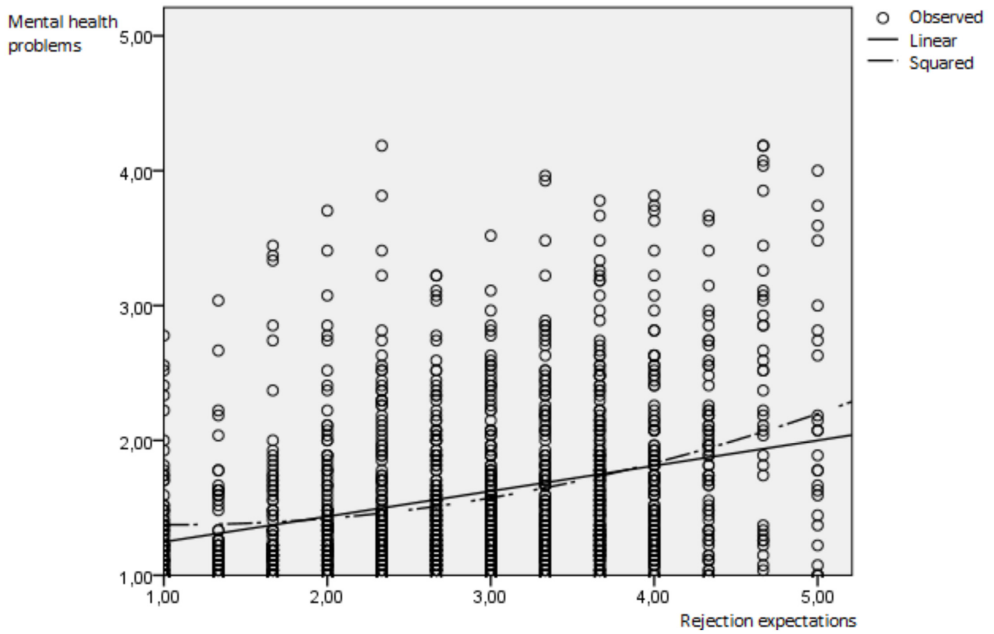


FIGURE 2 | Linear and convex prediction of mental health problems by rejection expectations. Observed = observed value, linear = linear line of parameter estimation, squared = squared (convex) line of parameter estimation.

found than expected by both predictors separately (compare **Figures 1, 2**). Note that when victimization squared and rejection expectations squared were used as individual predictors in the polynomial regression, only rejection expectations squared showed a significant prediction of mental health problems. It is possible that gay and bisexual men are overloaded by a high number of victimization events and especially by a high level of rejection expectations leading to a stronger increase in mental health problems. Another explanatory model is that gay and bisexual men with higher levels of mental health problems may overestimate their level of victimization and rejection expectations as found in individuals with depression due to their tendency for biased attention, processing, thoughts, and memory (Disner et al., 2011). Future research is needed to replicate the findings as well as to test possible explanatory models.

Furthermore, we did not find that discrepant values in victimization and rejection expectations predicted mental health problems at a relevant level. While a significant prediction was found when victimization was higher than rejection expectations, the size of the prediction was at an irrelevant level. This implicates that it is slightly adaptive for gay and bisexual men to have a level of rejection expectations that is higher than or corresponding to the level of experienced victimization. A possible explanation could be that rejection expectations help gay and bisexual men to process victimization. However, since this relationship was very low, we interpret it as not externally relevant.

Moreover, longitudinal and experimental data would be especially useful in determining the direction of the prediction between rejection expectations and mental health.

Limitations of the study include that a cross-sectional approach was used. It is therefore possible that the predictions are inversed, i.e., mental health problems predicting higher victimization. A further limitation was that the used scales had not been previously validated and that the Cronbach's alpha coefficient of rejection expectation was between questionable (0.60) and sufficient (0.70). However, a *post hoc* factorial analysis confirmed a one-factor solution for this scale. Thereby, factorial validity of the scale could be established. Nevertheless, type-II errors derived from this scale are still more likely in the current study and the correlations between rejection expectations and victimization as well as rejection expectations and mental health problems are likely to be underestimated.

CONCLUSION

This study provides the first evidence for a curvilinear (upward curving) relationship between victimization, rejection expectations, and mental health problems. It also replicates findings documenting a linear relationship between victimization, rejection expectations, and mental health problems. Furthermore, discrepancies in victimization and rejection expectations are not associated with mental health problems.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Burton, C. M., Marshal, M. P., Chisolm, D. J., Sucato, G. S., and Friedman, M. S. (2013). Sexual minority-related victimization as a mediator of mental health disparities in sexual minority youth: a longitudinal analysis. *J. Youth Adolesc.* 42, 394–402. doi: 10.1007/s10964-012-9901-5

Disner, S. G., Beevers, C. G., Haigh, E. A. P., and Beck, A. T. (2011). Neural mechanisms of the cognitive model of depression. *Nat. Rev. Neurosci.* 12, 467–477. doi: 10.1038/nrn3027

Eaton, N. R. (2014). Transdiagnostic psychopathology factors and sexual minority mental health: evidence of disparities and associations with minority stressors. *Psychol. Sex. Orientat. Gend. Divers.* 1, 244–254. doi: 10.1037/sgd0000048

Eldahan, A. I., Pachankis, J. E., Rendina, H. J., Ventuneac, A., Grov, C., and Parsons, J. T. (2016). Daily minority stress and affect among gay and bisexual men: a 30-day diary study. *J. Affect. Disord.* 190, 828–835. doi: 10.1016/j.jad.2015.10.066

Feinstein, B. A., Goldfried, M. R., and Davila, J. (2012). The relationship between experiences of discrimination and mental health among lesbians and gay men: an examination of internalized homonegativity and rejection sensitivity as potential mechanisms. *J. Consult. Clin. Psychol.* 80, 917–927. doi: 10.1037/a0029425

Franke, G. H. (2000). *Brief Symptom Inventory von L. R. Derogatis (Kurzform der SCL-90-R)– Deutsche Version*. Göttingen: Beltz Test GmbH.

Fredriksen-Goldsen, K. I., Simoni, J. M., Kim, H.-J., Lehavot, K., Walters, K. L., Yang, J., et al. (2014). The health equity promotion model: reconceptualization of lesbian, gay, bisexual, and transgender (LGBT) health disparities. *Am. J. Orthopsychiatry* 84, 653–663. doi: 10.1037/ort0000030

Frisell, T., Lichtenstein, P., Rahman, Q., and Långström, N. (2010). Psychiatric morbidity associated with same-sex sexual behaviour: influence of minority stress and familial factors. *Psychol. Med.* 40, 315–324. doi: 10.1017/S0033291709005996

Herek, G. M., and Berrill, K. T. (eds). (1992). *Hate Crimes: Confronting Violence against Lesbians and Gay Men*. Newbury Park, CA: Sage Publications, Inc.

King, M., Semlyen, J., Tai, S. S., Killaspy, H., Osborn, D., Popelyuk, D., et al. (2008). A systematic review of mental disorder, suicide, and deliberate self harm in lesbian, gay and bisexual people. *BMC Psychiatry* 8:70. doi: 10.1186/1471-244X-8-70

McLaughlin, K. A., Hatzenbuehler, M. L., Xuan, Z., and Conron, K. J. (2012). Disproportionate exposure to early-life adversity and sexual orientation disparities in psychiatric morbidity. *Child Abuse Negl.* 36, 645–655. doi: 10.1016/j.chiabu.2012.07.004

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological Bulletin* 129, 674–697. doi: 10.1037/0033-2909.129.5.674

Nathans, L. I., Oswald, F. I., and Nimon, K. (2012). Interpreting multiple linear regression: a guidebook of variable importance. *Pract. Assess. Res. Evaluat.* 17, 1–19.

Pachankis, J. E. (2015). A transdiagnostic minority stress treatment approach for gay and bisexual men's syndemic health conditions. *Arch. Sex. Behav.* 44, 1843–1860. doi: 10.1007/s10508-015-0480-x

Pachankis, J. E., Goldfried, M. R., and Ramrattan, M. E. (2008). Extension of the rejection sensitivity construct to the interpersonal functioning of gay men. *J. Consult. Clin. Psychol.* 76, 306–317. doi: 10.1037/0022-006X.76.2.306

Pachankis, J. E., Hatzenbuehler, M. L., and Starks, T. J. (2014a). The influence of structural stigma and rejection sensitivity on young sexual minority men's daily tobacco and alcohol use. *Soc. Sci. Med.* 103, 67–75. doi: 10.1016/j.socscimed.2013.10.005

Pachankis, J. E., Rendina, H. J., Restar, A., Ventuneac, A., Grov, C., Parsons, J. T., et al. (2014b). A minority stress—emotion regulation model of sexual compulsivity among highly sexually active gay and bisexual men. *Health Psychol.* 34, 829–840. doi: 10.1037/hea0000180

Sattler, F. A., Wagner, U., and Christiansen, H. (2016). Minority stress, group-level coping, social support, and mental health of German gay men. *PLoS ONE* 11:e0150562. doi: 10.1371/journal.pone.0150562

Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., and Heggestad, E. D. (2010). Polynomial regression with response surface analysis: a powerful approach for examining moderation and overcoming limitations of difference scores. *J. Bus. Psychol.* 25, 543–554. doi: 10.1007/s10869-010-9183-4

Conflict of Interest Statement: 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.

Copyright © 2017 Sattler and Christiansen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 16 October 2017
doi: 10.3389/fpsyg.2017.01761

Expectation Violation in Political Decision Making: A Psychological Case Study

Michael Öllinger[1,2], Karin Meissner[3,4], Albrecht von Müller[1,5] and Carlos Collado Seidel[6]*

[1] Parmenides Foundation, Pullach, Germany, [2] Department of Psychology, Ludwig Maximilian University of Munich, Munich, Germany, [3] Institute of Medical Psychology, Ludwig Maximilian University of Munich, Munich, Germany, [4] Division Integrative Health Promotion, University of Applied Sciences Coburg, Coburg, Germany, [5] Department of Philosophy, Ludwig Maximilian University of Munich, Munich, Germany, [6] Department for Modern and Contemporary History, Phillips University of Marburg, Marburg, Germany

Since the early Gestaltists there has been a strong interest in the question of how problem solvers get stuck in a mental impasse. A key idea is that the repeated activation of a successful strategy from the past results in a mental set ('Einstellung') which determines and constrains the option space to solve a problem. We propose that this phenomenon, which mostly was tested by fairly restricted experiments in the lab, could also be applied to more complex problem constellations and naturalistic decision making. We aim at scrutinizing and reconstructing how a mental set determines the misinterpretation of facts in the field of political decision making and leads in consequence to wrong expectations and an ill-defined problem representation. We will exemplify this psychological mechanism considering a historical example, namely the unexpected stabilization of the Franco regime at the end of World War II and its survival thereafter. A specific focus will be drawn to the significant observation that erroneous expectations were taken as the basis for decisions. This is congruent with the notion that in case of discrepancy between preconceived notions and new information, the former prevails over the new findings. Based on these findings, we suggest a theoretical model for expectation violation in political decision making and develop novel approaches for cognitive empirical research on the mechanisms of expectation violation and its maintenance in political decision making processes.

Keywords: problem-solving, expectancy, expectation violation, political-decision-making, mental set

OPEN ACCESS

Edited by:
Bernhard Hommel,
Leiden University, Netherlands

Reviewed by:
Lesley K. Fellows,
McGill University, Canada
Derrick L. Hassert,
Trinity Christian College, United States

****Correspondence:***
Michael Öllinger
michael.oellinger@parmenides-foundation.org

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *10 November 2016*
Accepted: *22 September 2017*
Published: *16 October 2017*

Citation:
Öllinger M, Meissner K, von Müller A and Collado Seidel C (2017) Expectation Violation in Political Decision Making: A Psychological Case Study. Front. Psychol. 8:1761. doi: 10.3389/fpsyg.2017.01761

FIXATION, MENTAL SET, AND EXPECTATION

In the year 1935, 4 years before the Second World War started, Karl Duncker published his seminal book "Zur Psychologie produktiven Denkens" (Translated to English in 1945: "On problem solving") (Duncker, 1935, 1945). In his book Duncker founded a theory which has been widely influencing the research of insight problem solving until today. Insight problems are characterized by having either no obvious or just step-wise solutions, but they have a sudden, unexpected, and unintended character. Usually, the solution requires a re-structuring of the problem elements or the assumptions that were imposed on the problem (Metcalfe and Wiebe, 1987; Ohlsson, 1992, 2011; Wegner, 2002; Öllinger and Knoblich, 2009). Duncker was fascinated by the question, what factors block the problem solving process and make people blind to insightful solutions.

In his famous candle problem he asked participants to fix a candle on the wall – given a matchbox, a box full of tacks and a candle. The problem proved to be extremely difficult. The solution proceeds as follows: empty the box and fix it with the tacks to the wall, then light the candle, put wax on the matchbox and glue the candle onto the box. Duncker argued the problem was difficult because participants fixated on the usual functions of the given objects. In this case, the given box needs to be used as a container, not as a ledge. This example demonstrates that prior knowledge imposes constraints on the utilization of objects. Similarly, Maier showed a few years earlier (Maier, 1931) that participants had problems using an object (e.g., a pair of pliers) as a weight for a pendulum.

This is part of the solution to the two-string-problem (see **Figure 1**). Participants were asked to tie two strings together which hung from the ceiling. The distance between the strings was too far apart to catch hold of both at once. The "insightful" solution to this problem is to use the pair of pliers as weight on one string and to swing it like a pendulum. Thus, both ends of the strings can be reached. Participants fixated on the usual function of the pair of pliers. Therefore, they had difficulties to use the objects' weight properly.

Luchins (1942) demonstrated that fixation was not only restricted to object properties, but could also be induced by the repeated activation of the same successful solution procedure. Luchins asked participants to solve various water-jug-problems. The objective was to fill a certain amount of water into one of three jars with different capacities. The capacities of the empty jars in the first experiment were: A (21 units), B (127), and C (3). The goal was to attain exactly 100 units by pouring the water from one to another. The solution of the problem is to fill water into B (127), then pour water from B (127) to A (21) = B (106) and finally twice from B (106) to C (2×3) = B (100). Luchins provided a sequence of analogous problems which could always be solved with exactly the same sequence. After a sequence of similar problems, a test problem was presented, which could be solved either by the usual sequence or in a much easier way [e.g., A(23), B(49), and C(3) goal state was 20]. The easy solution is to pour water from A to C. Almost two thirds of the participants who were trained with the difficult strategy were blind to the easy solution. They were caught in a mental set. Luchins also showed that even in the case of a problem which obviously could not be solved by the learned procedure, participants tried the usual strategy and re-applied it subsequently when a new problem was presented. It seemed fairly difficult to overcome the mental set and the associated expectations.

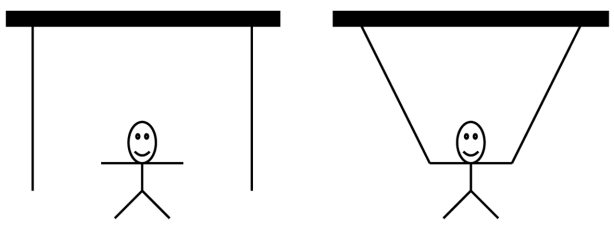


FIGURE 1 | Maier's two-string problem. Left: Initial state. Right: Goal state.

Lovett and Anderson (1996) proposed a computational model for mental set. It demonstrated that by increasing the weight of a procedure after each successful attempt the probability of its application will increase.

Öllinger et al. (2008) combined the concept of mental set with the domain of insight problem solving. They demonstrated that telling participants an insightful solution to a problem and repeating the same solution strategy several times inhibited the likelihood that participants applied a standard solution to problems – even if they did not require any insightful problem solving. As a result, this means insight blocked well-known prior knowledge strategies.

To conclude, we propose that fixation and mental set induce rigid behavior – firstly by exploiting prior knowledge and secondly by procedural and working memory activation. It seems conceivable to assume that both mechanisms influence the problem solvers' expectation. Fixation constrains the expectation about the utilization of an object. Mental set creates an expectation about the most promising and efficient strategy. By failures, expectations are violated and participants get stuck in an impasse or reluctantly repeat the wrong solution approach (Smith and Blankenship, 1989, 1991; Fedor et al., 2015; Öllinger et al., 2016).

Carnevale and Probst (1998) investigated the question what kind of mental sets were introduced by either social conflicts or social cooperation. In a first experimental group, they induced a conflict by giving the information that others will compete in a negotiation situation. In the second group, participants were informed that others want to cooperate with them. After mental sets were induced, the participants were asked to individually solve Duncker's candle problem. It turned out that participants in the conflict situation were significantly less likely to find the creative solution (empty the box and use it as a platform) than in the cooperation set. The authors emphasized that considering a situation as a conflict "promotes a freezing of knowledge" (p. 1301).

Bar-Tal et al. (1989) suggested that changing expectations could help to resolve a conflict mental set and convey it to a cooperation set. The authors stated that a social conflict is a cognitive schema. The schema is associated with knowledge and implications emanated by core beliefs.

Kruglanski (2013) postulated two processes. The first is the generation phase, which generates cognitive content. The second is the cognition validation phase, where a degree of confidence is mapped to the generated content. The first phase is crucial for potential mental set in our framework, since it selectively sets its focus on selected and biased pieces of information. In the second phase, persons test the generated information with stored evidence, whether the information is logically consistent or not.

Bar-Tal et al. (1989) detailed on potential processes which influence the generation process. They argued that parties have a need for closure, which means to stick to certain beliefs or maintain a particular belief as true and reject contrary ideas or alternative perspectives. To resolve a conflict, it is necessary to

alternate the cognitive schema. The expectation that a conflict will continue will not change the accessibility of the conflict schema.

For our line of argumentation this means that the violation of expectations will activate the conflict schema. The repeated activation of this mismatch will strengthen the conflict and at the same time strengthen the core believe that there is only one solution to the given problem – the person gets stuck in a mental impasse.

To sum up, there is evidence that decision making processes are negatively influenced by the repeated activation of an apparently successful solution strategy or apparently related information. Furthermore, successful strategies appear to increase the likelihood for a premature closure making blind for alternative solution approaches. Regarding political decision making, this may imply that political actors can find themselves in a conflict mode which prevents creative thinking necessary to solve a difficult problem.

In the following, we aim at contextualizing historically documented facts based on a comprehensive archival research (see Collado Seidel, 2016) with the purpose of demonstrating the relevance of mental sets in the domain of political decision making. We will exemplify this psychological mechanism considering the unexpected stabilization of the Franco regime at the end of World War II and after. Based on cognitive models, we aim to explain the persistence of a contradiction between expectations on the one hand and the rational perception of given facts pointing in the opposite direction on the other hand. A specific focus will be drawn to the significant observation that erroneous expectations were taken as the basis for decisions, showing that in case of discrepancy between preconceived notions and new information the former prevail over the new findings.

FRANCO AND THE EXPECTED POST-WAR ORDER: RECONSTRUCTION OF EXPECTATION VIOLATION IN A HISTORICAL CONTEXT

The Franco regime was considered an intrinsic part of the fascist European order during World War II (see Bowen, 2000; Collado Seidel, 2001, 2005, 2012). This circumstance was clearly perceived by the British and Americans. As a striking example, in 1940 the British ambassador to Madrid, Sir Samuel Hoare, wrote to a member of the Cabinet: "I have never seen so complete a control of the means of communication, press, propaganda, aviation, etc., as the Germans have here. Indeed, I go so far as to say that the Embassy and I are only existing here on German sufferance" (Hoare, 1946, p. 32).

Therefore, American and especially British political observers and decision makers, who had the leading role in the definition of Allied politics toward Spain, expected that with the crushing of German Nazism and Italian fascism, the Spanish dictator would inevitably fall as well. A radical change, hopefully by democratic forces, was expected and the conviction persisted that this outcome was merely a matter of time and the "problem Franco" would solve itself.

This conviction was consequently expressed by diplomats and politicians involved and was shared at the top level of the British and American governments as well (Hull et al., 1943; Hollis, 1944). It can be illustrated with one of Hoare's vivid appraisals, dated June 1943: "The Spanish tide is, in fact, running in our favor and, this being so, I should let it take its own course, and not attempt to force its pace. [...] It will, in my considered view, collapse all the sooner if we leave it to the Spaniards themselves to give it the *coup de grâce*. [...] The evidences available in Spain go to show that it will be a monarchist restoration, and that the restoration will be attempted between now and the end of the year" (Hoare, 1943). Oliver Harvey, the Principal Private Secretary to the British Foreign Secretary, put it straight shortly thereafter with his remark: "Damn Franco! We'll have him off his perch before we are done" (Harvey, 1943).

Remarkably, this wrong expectation was maintained despite the perception that Franco was even strengthening his power within Spain. Psychologically, one may assume that the responsible decision makers suffered from a pronounced mental set which drove their judgment.

EXPECTATION: THE FRANCO REGIME WILL NOT SURVIVE WORLD WORD II

The expectation of the Allied that the Franco Regime would not survive World War II was based on three mutually related assumptions:

(1) If the fascist reference system (Axis Powers) collapses, Spain will be destabilized.

(2) The opposition in Spain will exploit the weakness of the Franco regime and will establish a new regime along democratic structures; otherwise, revolutionary events will force the issue.

(3) The Spaniards are eager to get rid of Franco.

In anticipation of the results: All of these expectations were proven wrong and in the end the Franco regime proved stable and kept control until 1977, when a new democratic constitution was worked out and the first free elections were held in Spain.

In the last stages of World War II, however, the British and American governments started from the premise that Franco's Spain would not survive the collapse of the Axis Powers. Furthermore, observers – such as the British ambassador – were convinced that the dissatisfaction with the regime was growing continuously in all sectors of Spanish society. Ambassador Hoare professed moreover his belief that the vast majority of Spaniards, even the working class, favored the restoration of the Monarchy (Hoare and Greville, 1942; Hoare, 1944).

The American ambassador to Spain, Carlton Hayes, basically shared this view, though he, as the Americans in general, favored the establishment of a Republic. Furthermore, though he had serious doubts whether the Spaniards really preferred the Monarchy being restored, he firmly shared the conviction that the vast majority of Spaniards detested the Falange and that Franco would leave power either voluntarily or forcibly, giving way to a new regime (Hayes, 1944). In case of Franco's refusal to permit a

political transition or to introduce radical changes in the structure of his regime, the political observers expressed their conviction that the dictator would be forced to leave power, as ambassador Hayes put it in May 1943: "If Franco gets rid of the Falange in time (which I imagine he won't), he may be able to lead an evolution toward a more liberal government and to retain a place in it. Otherwise, he will be forcefully ousted along with the Falange" (Hayes, 1943).

The expectation of a sudden breakdown of the Franco regime became even more intense and bordered on certainty after the dismissal of the fascist dictator Mussolini in July 1943: Observers like Alan Hillgarth, the key person of British intelligence services in Spain, or George Kennan, special envoy of the US-State Department, were persuaded that the Italian events would shortly find a repetition in Spain (Hillgarth, 1943; Kennan, 1943).

In sum, there existed no doubts that Franco's end was at hand. The British and Americans maintained this conviction as shown exemplarily in a dispatch from the US-ambassador, dated September 1944: "The régime, as it is, can hardly survive the final outcome of the war. In a Europe, and a world, then turning more and more 'leftward,' Spain could not remain apart and insulated from such a universal current" (Hayes, 1944).

Contrary to these expectations, however, neither the destitution of Mussolini in July 1943 nor the landings of the Allied Forces in Normandy in June 1944, and not even the collapse of Nazism in May 1945 shook the Spanish regime. Nevertheless, the Western Allies perseverated on their view, as a British diplomat put it in October 1945, though in a somehow exasperated way in the face of the past experiences: "Franco's down-fall is only a matter of time, whether weeks, or months or years" (Garran, 1945b). Realizing that the expected changes did not occur, the political analysts seemed stunned and helpless and got stuck in an impasse: "There is a Spanish political reality entirely apart from the general European situation, i.e., Franco needs not fall with Hitler though undoubtedly Franco will fall if he does not change in time" (Bonsal, 1944).

PERCEPTIONS RUNNING CONTRARY TO THE EXPECTATIONS

This conviction professed by the British and Americans is especially surprising due to the circumstance that the same political observers perceived the weakness and disunity of the oppositional forces to compel a change and on the contrary, the political system showed no signs of weakening. The monarchists as the most promising group just launched weak attempts by presenting writings in which they urged in favor of the restoration of the Monarchy. In addition, they were profoundly divided in view of the pursued aims after the downfall of Franco. This was understood very clearly: "The strong sectionalism and individualism within Spain appear to foster sectional and individual political aims which transcend national political aims, with consequent lack of any unifying national program or organization for the opposition as a whole or even for any considerable part of it" (Hayes, 1944). The same basic problem was seen in the case of the republican movement, whose only common factor was to overthrow the existing regime. Even the communist guerrilla, as the best organized opposition group, showed itself unable to force a shift and to provoke a general uprising, as demonstrated by the failure of the incursion of some thousand guerrilla fighters in the Pyrenees in autumn 1944.

In view of these developments, as early as February 1944, the British general staff reached the conclusion that no existing political or military group was in a position to oust Franco (Joint Planning Staff, 1944). Half a year later, a British political analyst summarized the situation in Spain in the same way by stating: "Contrary to normal expectations General Franco's position is undoubtedly stronger in Spain today than it had been at any time during the past few years" (Roberts, 1944). Even after the Hitler regime collapsed, Franco showed no tendency to change his politics. He had apparently even consolidated his position, as assessed by a British diplomat: "Franco is at present more firmly established in power than ever, in spite of the defeat of his fellow dictators" (Garran, 1945a).

Besides the observation that the opposition was deeply divided, British and American analysts realized quite early the presumably most ponderous reason for the hesitant attitude of the Monarchists as the most promising opposition group: the fears of provoking a communist revolution and the return of uncontrolled violence and chaos in case of a sudden political change and in particular by destabilizing the system in trying to overthrow Franco forcibly. The profiteers of the outcome of the Spanish Civil War did not want to endanger their own privileged economic and social position for the sake of bringing back the King. In the end, they felt quite comfortable with the prevailing situation. Thus, Hoare remarked with an amazed undertone he had "never known so many professed monarchists who didn't really want a king" (quoted in Hayes, 1946, p. 269).

Furthermore, the Allied perceived that Franco not only presented himself as the sole person able to prevent the country from plunging into chaos and anarchy. As stated by British diplomats, in view of the ostracism practiced by the British and American governments the dictator managed to present the criticism against his regime and his person as an attack on the Spanish nation, achieving broad public support (see Portero, 1989, p. 217; Collado Seidel, 2015, pp. 178f.).

DECISION MAKERS IN CONFLICT

The fear of provoking a renewed civil war hindered not only the Monarchists in their attempts to overthrow the Franco regime. The Allied, and especially the British policy makers as spokesmen of the common politics toward Spain, were as well discouraged from pursuing a more straightforward policy accompanied by the imposition of effective economic sanctions such as an oil-embargo for the same reason. Churchill put it straight by remarking toward a more receptive Foreign Secretary: "What you are proposing to do is little less than stirring up a revolution in Spain. You begin with oil: you will quickly end in blood" (Churchill, 1944b).

But it was not just the memory of the atrocious civil war, which hindered the enforcement of a rather stiff attitude toward Franco. Above all, Churchill had in mind the Soviet Union and its alleged interests in provoking a revolution and a communist takeover on the Iberian Peninsula in mind. The chaotic and troublesome situation in Europe after the war led Churchill to avoid the embarkment toward new political adventures (Churchill, 1944a).

WRONG ASSUMPTIONS AND MISPERCEPTIONS

Our historical analysis allows us to assess the information disregarded by the Allied political decision makers, which prevented them to build up a proper representation of the problem. We reconstruct the facts that were available but were not integrated in a proper problem representation.

First, the assumption that the successful elimination of Hitler's Nazism and Mussolini's fascism would result in the downfall of the Franco regime, due to the consideration that it constituted an intrinsic part of the fascist European order, was the main misconception and led to dramatic misjudgements. Not least, the political actors were influenced by the events which led to the destitution of Mussolini. This resulted in a "sit and wait" attitude in the expectation that the problem solves itself.

Second, it was overestimated that the Spanish monarchical and republican opposition would exploit the doom of fascism bundling their forces with the aim of overturning the regime. On the contrary, the opposition was weak and disunited.

Third, while it can certainly be assessed that the dictator and the dictatorship were by no means popular, it proved wrong that the Spaniards were eager to get rid of the regime. This was not just the case in view of the monarchical opposition, which preferred not to run risks, which could endanger their own privileged position, but also for the vast majority of the population who was afraid of chaos and a renewed civil war, as shown by the lack of support the activities of the guerrilla received.

Furthermore, the British and American decision makers were trapped in a conflict mode regarding the putative reactions of the Soviet Union as a competing player: the risk of a revolutionary insurrection and a renewed civil war, which could lead to negative consequences for themselves, resulted in a narrow-minded assessment of the situation that impeded considering alternative procedures as postulated by Carnevale and Probst (1998).

To sum up: British and American observers and decision makers gained and discussed the information which proved the basic assumption of an imminent end of the Franco dictatorship wrong. Despite seriously analyzing the information, they adhered firmly to their original considerations. The search for alternatives may further have been hindered by a conflict mode resulting from the perception of the Soviet Union as a competing player. As a result, the political key players were stuck in an impasse, which they were unable to solve, and which contributed, in a historical perspective and against all expectations, to the outcome of the survival of the Franco regime after World War II until 1977.

ANALOGIES BETWEEN IMPASSE IN PROBLEM SOLVING AND HISTORICAL DEADLOCKS

This article follows the credo of the famous Gestaltist and social psychologist Kurt Lewin, stating that "there is nothing as practical as a good theory" (Tolman, 1996, p. 31). We now aim at bridging the gap between basic research on problem solving and the complex and applied historical situation presented above.

Expectations in problem solving, as described above, are driven either by prior knowledge or the repeated activation of a successful solution procedure (Lovett and Anderson, 1996; Ohlsson, 2011). Recently, Öllinger et al. (2016) demonstrated that even the repeated activation of an inappropriate solution strategy led to fixation on this strategy, although participants received consequent negative feedback.

The reason might be that within the ill-defined search space there were no other possibilities. This could nicely be demonstrated by the Nine-Dot problem (see **Figure 2**). The task is to connect the given nine dots by four straight and connected lines. It is not allowed to lift the pencil or to retrace a line (Maier, 1930; Scheerer, 1963). The problem proved to be extremely difficult and reluctant to clues or hints (Lung and Dominowski, 1985; Chronicle et al., 2001). Even when people were told that the solution implies to draw lines outside the given nine dots, most of the participants failed (Weisberg and Alba, 1981). Indeed, the majority of the naïve participants tried to solve the problem within the boundaries of the given nine dots. Öllinger et al. (2014) suggested that the Nine-Dot problem requires not only to relax the constraint, but also to have an appropriate strategy that helps to restrict the new and even larger search space, since drawing lines to non-dot points increases the search space exponentially.

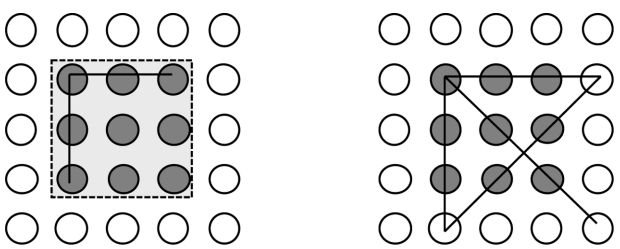


FIGURE 2 | Left: The constrained Nine-Dot problem. Right: A possible solution of the problem. In the beginning the Nine-Dot problem seems very simple and straightforward. Initially, problem solvers expect that the given nine dots need to be connected by dot-to-dot connecting lines and the self-imposed assumption that lines should stay within the boundaries of the given nine dots.

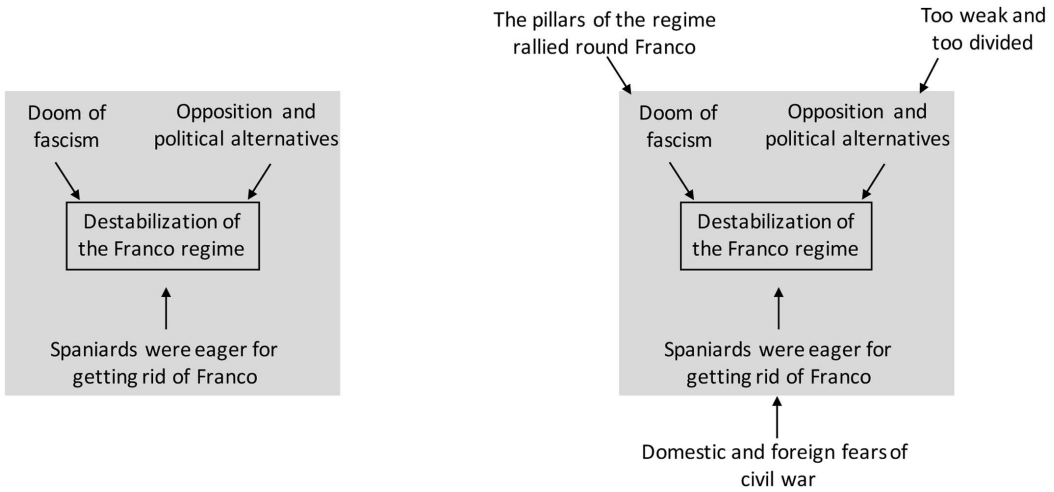


FIGURE 3 | Left: Over-constrained problem representation. Arrows indicate a positive influence. Right: Actual problem representation.

In analogy to our historical problem constellation, it seems conceivable that political leaders selected strategies from a too narrow and constrained problem space (see **Figure 3**). The solution seemed obvious and was only a matter of time. The historical documents provide unequivocal evidence that political observers persisted on their assessments although the actual development was fairly different. Retrospectively, these misjudgements seem inexplicable, because most facts were already available from the beginning and were perceived accordingly. Considering these facts would have resulted in a more reliable expectation. This line of argumentation builds on the work of Klein (1999, 2008), who provided a model on natural decision making. Klein introduced a step-wise model which relies on a recognition-primed decision process. Initially, experts utilize familiar and already approved solution strategies. In more detail, incidents activate prior knowledge and decision makers evaluate, whether they already met such a situation. If so, they execute the according actions. Depending on how complicated the current incident is, further steps (mental simulations and evaluation) are necessary, until an appropriate solution strategy is selected.

Tetlock (1996, 1999) showed that providing counterfactual facts to experts about historical events leads to clearly biased problem representations. The representations were determined by the decision makers' attitudes. The results revealed that experts neutralize dissonant data and preserve confidence in their prior assessments by resorting to a complex battery of belief-system defense. Tetlock showed that the results of what-if constructions are determined by the persons' ideological world view. He argues that experts confronted with counterfactual evidences attempt to reduce the cognitive dissonance (Festinger, 1962) by ignoring or biasing the given evidences.

We suggest that our approach goes beyond a recognition account. We identified mental set as a driving force for ill-defined and biased problem representations that mainly drives the selection of decision making strategies.

A PRELIMINARY MODEL ON POLITICAL DECISION MAKING BASED ON INSIGHT PROBLEM SOLVING

In this section we outline a model on political decision making, which is based on cognitive processes stemming from the domain of insight problem solving.

As **Figure 4** depicts, experts in the field acquired a large corpus of domain related knowledge by their profession. However, this profound knowledge is affected by attitudes, interests, school of thoughts, prior experience, current political tendencies, the contemporary discourse, political systems, and general political opinions (Tetlock, 1996, 1999). The application of the biased knowledge to a new political situation or an unknown counterfactual scenario (Bar-Tal et al., 1989; Tetlock, 1999) may result in a mental set which over-constrains the search space. As a result, familiar and well-known, but insufficient strategies are applied. Unfortunately, those strategies will not necessarily solve the problem (Öllinger et al., 2014). We assume that political experts have high confidence in the reliability and validity of their knowledge. Consequently, they will probably repeatedly activate the maladaptive solution strategy and see no need to change the solution process. This behavior will induce a mental set. The mental set prevents the decision makers' realization of expectation violation. Without the realization there is no drive to re-structure the self-imposed constrains. Realizing the violation of expectation will lead to a revised decision making process which relaxes the self-imposed constrains. As a result, the search space for a solution will be extended and a potential solution is accessible.

On a behavioral level, we assume that mental set is detectable by the used words (Cohn et al., 2004; Chung and Pennebaker, 2007; Pennebaker et al., 2015) (see our suggestions in the next section), the solution attempts, the thinking styles and the stream of the evaluative process.

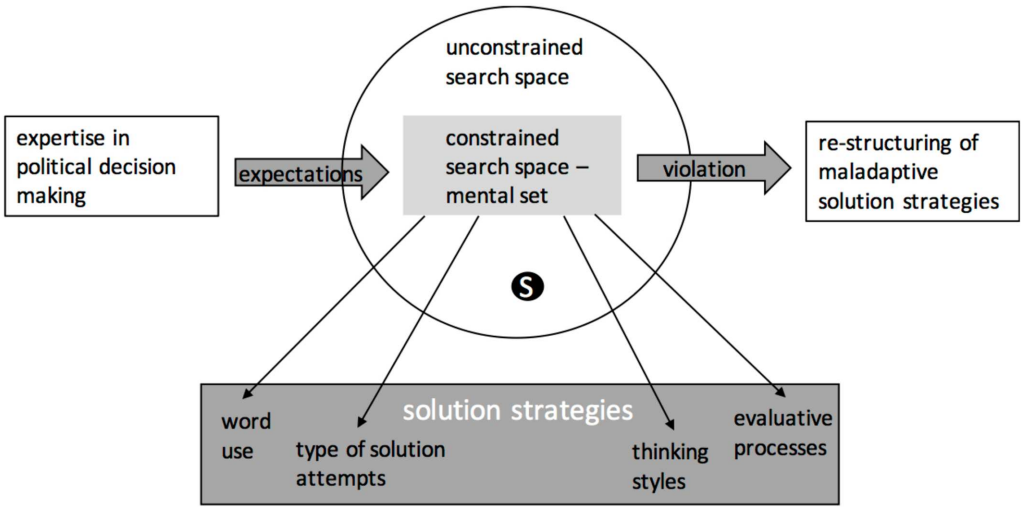


FIGURE 4 | Model of expectation violation in political decision making. Expertise creates expectations, which result in a constrained search space. If the solution (S) lies outside the constrained search space, maladaptive solution strategies will occur. The realization of expectation violation is the key ingredient to change the solution strategy. Behaviorally, mental set manifests by the word use, the type of solution attempts, thinking styles, and evaluative processes.

TESTABILITY OF OUR ASSUMPTION

To convey our case study and our model proposal into a testable research program, we derived the following hypotheses.

Word Count Analysis of the Diplomatic Documentation

We aim at analyzing the large documentary evidence, on which our study is based, by means of quantitative methods. Pennebaker et al. (2001, 2003), Tausczik and Pennebaker (2010), and Pennebaker (2011) analyzed the frequency of words and made predictions about personality and thinking styles (e.g., function words or the frequency of used pronouns). We postulate that the documented assessments of Western politicians will show similar linguistic structures. This would demonstrate that an ideological mental set determines the search space for potential solutions to the discussed problems in Spain. According to Klein's (2008) interpretation (see above), the situation was recognized as familiar and consequently, familiar solution strategies were applied.

Field Studies

We plan to address the question how expertise and training in a particular political or historical domain (see Tetlock, 1999) will influence the decision making process. We are following Klein's naturalistic decision making account (Klein, 1999, 2008) and plan to test experts from the field. We propose to confront them with fictive and complex historical and political scenarios, which the experts should evaluate. Then the experts will be asked to make predictions about the future developments of the scenarios. We hypothesize that the evaluation process will rather be determined by the specific expertise of the decision maker than by motivational factors like reducing cognitive dissonance (Tetlock, 1996).

Laboratory Studies

Very recently, Salvi et al. (2016) demonstrated at a behavioral level that liberals solved problems significantly more often with insight than conservative people. The authors argue that both groups have different cognitive styles. In general, liberals are more flexible whereas conservatives are more rigid and prefer clear answers. Therefore, conservatives solve problems more analytically and in a step-by-step manner, whereas liberals solve the problems non-step-wise by insight.

Wiley (1998) showed that baseball experts, who were asked to use words from their domain knowledge in an unusually context, revealed a significant mental set. The mental set prevents the solution of problems, which refers to the domain of expertise, but was used in a more remote context.

Taking the evidences from both studies, we propose to test historical and political experts with problems that either require domain related knowledge or not. We suggest that expertise will inhibit innovative solutions, when decision makers are asked to find unusual solutions.

Moreover, it would be helpful to split the participants into two extreme groups. The criterion would be the participants' cognitive styles. This would allow investigating the interaction between expertise and cognitive style and its impact on the solution of insight and non-insight problem solving as well as on political decision making.

CONCLUSION AND PERSPECTIVES

We demonstrated that insight problems violate the expectations of problem solvers by self-imposed constraints and mental set. We extrapolated these findings to more complex problems such as political decision making processes. By analyzing the considerations about Franco's Spain we showed that in fact

decision makers' limited presuppositions biased the search space which consequently led to inappropriate expectations. Interestingly, the biased representation was not even updated although there was evidence which contradicted the initial expectations.

Given these findings, we conclude that expectation violation could be viewed as a general process that plays an eminent role even for complex problem solving. In our understanding both insight problem solving and political decision making could benefit from a better understanding and a more thorough and detailed analysis of expectations. A potential indication follows directly from this conclusion. We suggest that it might be worthwhile to monitor the problem solving process in order to detect the violation of expectation. The feedback of such a monitoring process might help the problem solvers to update, to elaborate or to restructure the search space and the related expectations.

Based on these findings we suggest further research in a twofold direction: on the one side, it would be helpful to enhance the described analogies between a determined historical setting and the psychological models by means of analyzing further unexpected historical developments, which led to the emergence of problems in view of a correct political perception of the events. This will provide the data base for attaining verifiable conclusions. On the other side, problem solving strategies in a political context should also be followed up in experimental studies. For example, it would be interesting to disentangle the different brain areas involved in the persistence of political expectations when the truth of these expectations is challenged by contradictory knowledge or experience.

AUTHOR CONTRIBUTIONS

MÖ contributed the psychological part. CCS provided the historiographical part. KM worked on the line of argumentation. AvM provided the philosophical and political background.

FUNDING

KM received support by the Schweizer-Arau Foundation and the Theophrastus Foundation.

ACKNOWLEDGMENT

Parts of the work were presented at the Neurohistory Workshop of the Rachel Carson Center for Environment and Society at the LMU Munich, 2011.

REFERENCES

Bar-Tal, D., Kruglanski, A. W., and Klar, Y. (1989). Conflict termination: an epistemological analysis of international cases. *Polit. Psychol.* 10, 233–255. doi: 10.2307/3791646

Bonsal, P. (1944). *Diary entry, November 2, 1944,* Library of Congress, Washington (LC), Bonsal Papers.

Bowen, W. H. (2000). *Spaniards and Nazi Germany. Collaboration in the New Order*. Columbia: University of Missouri Press.

Carnevale, P. J., and Probst, T. M. (1998). Social values and social conflict in creative problem solving and categorization. *J. Pers. Soc. Psychol.* 74, 1300–1309. doi: 10.1037/0022-3514.74.5.1300

Chronicle, E. P., Ormerod, T. C., and MacGregor, J. N. (2001). When insight just won't come: the failure of visual cues in the nine-dot problem. *Q. J. Exp. Psychol.* 54 A, 903–919. doi: 10.1080/713755996

Chung, C., and Pennebaker, J. W. (2007). "The psychological functions of function words," *Social Communication*, ed. K. Fiedler (New York, NY: Psychology Press), 343–359.

Churchill, W. (1944a). *To the US-President, Telegram*. Chartwell Trust Papers (CHAR), 20, 165, 102, Cambridge: Churchill Archives Centre (CAC).

Churchill, W. (1944b). *To the Foreign Secretary, minute, November 10, 1944, Records of the Prime Minister's Office: Prime Minister's Correspondence Papers, 1945-1951 (PREM 8)*. London: Public Record Office/National Archives, 106.

Cohn, M. A., Mehl, M. R., and Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychol. Sci.* 15, 687–693. doi: 10.1111/j.0956-7976.2004.00741.x

Collado Seidel, C. (2001). *Angst vor dem "Vierten Reich". Die Alliierten und die Ausschaltung des deutschen Einflusses in Spanien,* 1944-1958. Paderborn: Schöningh.

Collado Seidel, C. (2005). *España, Refugio Nazi*. Madrid: Temas de Hoy.

Collado Seidel, C. (2012). "España en la segunda guerra mundial. La 'hábil prudencia' de un 'neutral," in *En el combate por la Historia. La república, la guerra civil, el franquismo*, ed. A. Viñas (Barcelona: Pasado & Presente), 593–612.

Collado Seidel, C. (2015). *Franco. General, Diktator, Mythos*. Stuttgart: Kohlhammer.

Collado Seidel, C. (2016). *El telegrama que salvó a Franco: Londres, Washington y la cuestión del Régimen (1942–1945)*. Barcelona: Crítica.

Duncker, K. (1935). *Zur Psychologie des Produktiven Denkens. Berlin-Heidelberg-*. New York, NY: Springer-Verlag.

Duncker, K. (1945). *On Problem-Solving*, Vol. 58. Washington, DC: American Psychological Association INC.

Fedor, A., Szathmáry, E., and Öllinger, M. (2015). Problem solving stages in the five square problem. *Front. Psychol.* 6:1050. doi: 10.3389/fpsyg.2015.01050

Festinger, L. (1962). *A Theory of Cognitive Dissonance*, Vol. 2. Redwood: Stanford university press.

Garran, P. (1945a). *Minute, July 1, 1945, Records of the Foreign Office: General Correspondence: Political (FO 371), 49612, Z 8559*, London: Public Record Office/National Archives.

Garran, P. (1945b). *Minute, October 9, 1945, Records of the Foreign Office: General Correspondence: Political (FO 371), 49613, Z 11632*. London: Public Record Office/National Archives.

Harvey, O. (1943). *Diary Entry, Hayes Papers (HP), 56399*. London: British Library (BL)

Hayes, C. (1943). To the US-President, Letter, May 3, 1943, Franklin Delano Roosevelt Library, Hyde Park, N.Y. (FDRL), Franklin D. Roosevelt Papers as President, President's Secretary's File (PSF) 50.

Hayes, C. (1944). To the Secretary of State, dispatch, September 26, 1944, National Archives, Washington (NA), Register Group (RG) 59, 852.00/9–2644.

Hayes, C. H. J. (1946). *Wartime Mission in Spain, 1942-1945*. New York, NY: Macmillan.

Hillgarth, A. (1943). *To the Admiralty, telegram, June 16 1943, Records of the War Cabinet and Cabinet Office: Cabinet Office, Special Secret Information Centre: Files (CAB 121)*, London: Public Record Office/National Archives, 512.

Hoare, S. (1943). *To the Foreign Secretary, Letter, June 23, 1943*, Templewood Papers (Tem.), C. VII, 5. Cambridge: Cambridge University Library.

Hoare, S. (1944). *To the Foreign Secretary, March 25, 1944*, Templewood Papers (Tem.), XIII, 24. Cambridge: Cambridge University Library.

Hoare, S. (1946). *Ambassador on Special Mission*. London: Collins.

Hoare, S., and Greville, M. (1942). *Letter, May 8, 1942, Cambridge University Library (CU),* Templewood Papers (Tem.), XIII, 4. Cambridge: Cambridge University Library.

Hollis, L. (1944). *To the Prime Minister, minute, February 10, 1944, Records of the War Cabinet and Cabinet Office: Cabinet Office, Minister of Defence Secretariat: Records (CAB 120)*. London: Public Record Office/National Archives, 692.

Hull, C., Butterworth, W., and Walser. (1943). *Memorandum, May 26, 1943, Cordell Hull Papers (PCH)*. Washington, DC: Library of Congress, 86.

Joint Planning Staff. (1944). *Memorandum, J.P. (43) 444 (S) (Final), February 7, 1944, Records of the Prime Minister's Office: Operations Papers, 1938-1946 (PREM 3), 405, 5*. London: Public Record Office/National Archives.

Kennan, G. (1943). *To the State Department, Telegram, September 20, 1943, Columbia University*, Papers of Carlton J.H. Hayes (CHP), New York, NY: Rare Book & Manuscript Library, 4.

Klein, G. (2008). Naturalistic decision making. *Hum. Factors* 50, 456–460. doi: 10.1518/001872008X288385

Klein, G. A. (1999). *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Kruglanski, A. W. (2013). *Lay Epistemics and Human Knowledge: Cognitive and Motivational Bases*. Berlin: Springer.

Lovett, M. C., and Anderson, J. R. (1996). History of success and current context in problem solving: combined influences on operator selection. *Cognit. Psychol.* 31, 168–217. doi: 10.1006/cogp.1996.0016

Luchins, A. S. (1942). Mechanization in problem solving–the effect of Einstellung. *Psychol. Monogr.* 54, 1–95. doi: 10.1037/h0093502

Lung, C.-T., and Dominowski, R. L. (1985). Effects of strategy instructions and practice on nine-dot problem solving. *J. Exp. Psychol.* 11, 804–811. doi: 10.1037/0278-7393.11.1-4.804

Maier, N. R. F. (1930). Reasoning in humans. I. On direction. *J. Comp. Psychol.* 10, 115–143. doi: 10.1037/h0073232

Maier, N. R. F. (1931). Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *J. Comp. Psychol.* 12, 181–194. doi: 10.1037/h0071361

Metcalfe, J., and Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Mem. Cognit.* 15, 238–246. doi: 10.3758/BF03197722

Ohlsson, S. (1992). "Information-processing explanations of insight and related phenomena," in *Advances in the Psychology of Thinking*, eds M. Keane and K. Gilhooly (London: Harvester-Wheatsheaf), 1–44.

Ohlsson, S. (2011). *Deep Learning: How the Mind Overrides Experience*. Cambridge: Cambridge University Press.

Öllinger, M., Fedor, A., Brodt, S., and Szathmáry, E. (2016). Insight into the ten-penny problem: guiding search by constraints and maximization. *Psychol. Res.* 8, 925–938. doi: 10.1007/s00426-016-0800-3

Öllinger, M., Jones, G., and Knoblich, G. (2008). Investigating the effect of mental set on insight problem solving. *Exp. Psychol.* 55, 269–282.

Öllinger, M., Jones, G., and Knoblich, G. (2014). The dynamics of search, impasse, and representational change provide a coherent explanation of difficulty in the nine-dot problem. *Psychol. Res.* 78, 266–275. doi: 10.1007/s00426-013-0494-8

Öllinger, M., and Knoblich, G. (2009). "Psychological research on insight problem solving," in *Recasting Reality*, eds H. Atmanspacher, and H. Primas (Berlin-Heidelberg: Springer), 275–300. doi: 10.1017/CBO9780511780295

Pennebaker, J. (2011). *The Secret Life of Pronouns: How Our Words Reflect Who We Are*. New York, NY: Bloomsbury.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi: 10.1027/1618-3169.55.4.269

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates. doi: 10.1007/s00426-013-0494-8

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annu. Rev. Psychol.* 54, 547–577.

Portero, F. (1989). *Franco Aislado. La Cuestión Española (1945-1950)*. Madrid: Aguilar.

Roberts, F. (1944). *Minute, October 7, 1944, Records of the Foreign Office: General Correspondence: Political (FO 371), 39677, C 13520*. London: Public Record Office/National Archives.

Salvi, C., Cristofori, I., Grafman, J., and Beeman, M. (2016). The politics of insight. *Q. J. Exp. Psychol.* 69, 1064–1072. doi: 10.1080/17470218.2015.1136338

Scheerer, M. (1963). Problem-Solving. *Sci. Am.* 208, 118–128. doi: 10.1146/annurev.psych.54.101601.145041

Smith, S. M., and Blankenship, S. E. (1989). Incubation effects. *Bull. Psychon. Soc.* 27, 311–314.

Smith, S. M., and Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem solving. *Am. J. Psychol.* 104, 61–87.

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54. doi: 10.1080/17470218.2015.1136338

Tetlock, P. E. (1996). *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton, NJ: Princeton University Press. doi: 10.1038/scientificamerican0463-118

Tetlock, P. E. (1999). Theory-driven reasoning about plausible pasts and probable futures in world politics: are we prisoners of our preconceptions? *Am. J. Pol. Sci.* 43, 335–366. doi: 10.3758/BF03334612

Tolman, C. W. (1996). *Problems of Theoretical Psychology*, Vol. 6. North York: Captus Press. doi: 10.2307/1422851

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press. doi: 10.1177/0261927X09351676

Weisberg, R. W., and Alba, J. W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. *J. Exp. Psychol. Gen.* 110, 169–192.

Wiley, J. (1998). Expertise as mental set: the effects of domain knowledge in creative problem solving. *Mem. Cognit.* 26, 716–730. doi: 10.2307/299 1798

Conflict of Interest Statement: 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.

Copyright © 2017 Öllinger, Meissner, von Müller and Collado Seidel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 27 December 2016
doi: 10.3389/fpsyg.2016.02024

Three Ways That Non-associative Knowledge May Affect Associative Learning Processes

Anna Thorwart[1]* *and Evan J. Livesey*[2]

[1] Department of Psychology, Philipps-Universität Marburg, Marburg, Germany, [2] School of Psychology, The University of Sydney, Sydney, NSW, Australia

Associative learning theories offer one account of the way animals and humans assess the relationship between events and adapt their behavior according to resulting expectations. They assume knowledge about event relations is represented in associative networks, which consist of mental representations of cues and outcomes and the associative links that connect them. However, in human causal and contingency learning, many researchers have found that variance in standard learning effects is controlled by "non-associative" factors that are not easily captured by associative models. This has given rise to accounts of learning based on higher-order cognitive processes, some of which reject altogether the notion that humans learn in the manner described by associative networks. Despite the renewed focus on this debate in recent years, few efforts have been made to consider how the operations of associative networks and other cognitive operations could potentially interact in the course of learning. This paper thus explores possible ways in which non-associative knowledge may affect associative learning processes: (1) via changes to stimulus representations, (2) via changes to the translation of the associative expectation into behavior (3) via a shared source of expectation of the outcome that is sensitive to both the strength of associative retrieval and evaluation from non-associative influences.

Keywords: associative learning, causal learning, expectation, prediction error, blocking

OPEN ACCESS

Edited by:
Hannes Ruge,
Dresden University of Technology, Germany

Reviewed by:
David Luque,
University of New South Wales, Australia
Miguel A. Vadillo,
King's College London, UK

***Correspondence:**
Anna Thorwart
anna.thorwart@staff.uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 October 2016*
Accepted: *13 December 2016*
Published: *27 December 2016*

Citation:
Thorwart A and Livesey EJ (2016) Three Ways That Non-associative Knowledge May Affect Associative Learning Processes. Front. Psychol. 7:2024. doi: 10.3389/fpsyg.2016.02024

INTRODUCTION

Associative theories of learning offer a powerful account of the way animals and humans assess the relationship between events and generate expectations about the future. They assume that we reflect our knowledge about the predictive relationships between events in associative networks, which consist of mental representations of these events and the associations that link them. These events could be predictive cues and subsequent outcomes in the case of Pavlovian learning or actions associated with antecedents and consequences in the case of instrumental learning. Through observing the co-occurrence of cues and outcomes, an individual learns the associations between them in such a way that the presence of a predictive cue brings to mind the outcome and thus informs subsequent behavior by generating an outcome expectation.

Associative accounts have been applied to many widely replicated learning phenomena. The focus of theory development over the last 50 years has included explaining how simultaneously

presented cues might compete for association, how selective attention affects and is affected by learning, and explaining how association formation could be a simple but effective means of tracking statistical contingencies between events rather than merely tracking their temporal coincidence. Not surprisingly, many associative learning models provide detailed and compelling explanations for these phenomena. Central to many of these explanations is the notion of *prediction error*, the discrepancy between what the associative system predicts will happen next and what is then actually experienced. The prediction error thus captures an experienced violation of expectations and we return to this concept and its widespread use in associative learning theory later. (Note also, we will use the terms expectation and prediction interchangeably). The term expectancy will usually refer to the explicit judgement of outcome expectations.

Despite the relative success that associative learning models have enjoyed in both explaining and predicting phenomena observed in conditioning and contingency learning studies, there are clearly also many influences on learning that associative models simply do not capture. In this article, we review some of these influences and speculate on how they might influence the operations of an associative learning system, assuming that such a system forms a core part of human learning and cognition more generally. A number of theoretical and empirical papers published in the last decade have approached this question and reviewed relevant literature in causal and contingency learning (e.g., Shanks, 2007; Mitchell et al., 2009; Boddez et al., 2014). The current paper does not intend to systematically and comprehensively review the same body of literature or provide a critique of the theoretical views proposed by these authors. Instead, our focus will be on outlining and discussing ways non-associative knowledge might influence the operations of an associative learning system without changing its fundamental principles. In doing so, we hope to provide a means of evaluating the contribution of theories based on associative networks to explaining complex behavior more broadly. We would argue that this is particularly relevant to human associative learning, where influences on behavior are clearly more complex than formal associative models can explain in isolation but where there is still support for the existence of association formation mechanisms. Some of the traditional sources of evidence have failed to convince all theorists that it is necessary to posit association formation as being mechanistically distinct from inferential reasoning or higher order cognition in general. For instance, the notion that associative learning can occur in the absence of awareness is still as contentious as ever (see Goujon et al., 2015; Colagiuri and Livesey, 2016; Vadillo et al., 2016 for a recent iteration of this debate concerning implicit learning in visual search). Nevertheless, a number of results (e.g., Morís et al., 2014; Perruchet, 2015; Cobos et al., 2016) suggest that associative learning mechanisms are separable from other cognitive sources of expectation in at least some circumstances and could represent the operation of an independent system. This possibility is certainly plausible enough to warrant a more in-depth consideration of how associative and non-associative sources of prediction might interact.

EXPECTANCY AND JUDGMENT IN HUMAN CAUSAL LEARNING

Studies of causal and contingency judgements are concerned with the way humans make explicit assessments of the predictive and causal relationships between events. These events may consist of a particular outcome in a fictitious scenario, such as an allergic reaction suffered by a patient, and the cues that may cause or predict that outcome, e.g., foods eaten by the patient. Participants will receive on a trial-by-trial basis information about what the patient has eaten. They are then requested to make a choice between different possible allergic responses that the patient might experience, e.g., "no allergic response," "rash," and "fever," to indicate their expectation, and receive feedback whether their expectation was correct and which symptoms the patient actually suffered after eating these foods. In a test phase after several trials of training, participants might be additionally asked to rate the relationship of certain foods with a certain allergic symptom on a scale from "not predictive/causal" to "highly predictive/causal."

One line of research has been to relate this kind of causal learning to classical conditioning and by this to associative accounts of learning. In addition to the apparent parallels in both the procedure and the content of learning – participants learn to predict future events based on their relationships with preceding events – many behavioral effects can be observed in both classical conditioning and human causal learning paradigms. However, when researchers started to investigate factors controlling learning in these kinds of procedures, critical factors quickly emerged that were not easily captured by associative models of learning, factors relating to "non-associative" knowledge relevant to the learning situation (see De Houwer, 2009).

ASSOCIATIVE AND NON-ASSOCIATIVE KNOWLEDGE: A WORKING DEFINITION

For the purpose of this paper, we will define *associative* knowledge as knowledge that can be derived merely from the statistical relationships among the relevant cues and outcomes. All knowledge that goes beyond this is then seen as *non-associative*. This concerns both the way this information is obtained and the content of the information. Non-associative knowledge includes information given verbally (i.e., by instruction, other people's accounts, prior semantic knowledge), relational information inferred from the co-occurrence of other, separately presented cues and outcomes, (e.g., whether other outcomes are predicted reliably by other cues), or information implied by other aspects of the experimental procedure (e.g., spatial position of cues and outcome on the screen, the format of the test question). We also regard information about properties of the associative links other than the statistical contingency, like their causal nature or the additivity of their effects, as non-associative, as well as information about properties of the cues and outcomes (e.g.,

whether the outcome is binary vs. continuous, and if continuous, whether it is observed at a maximal or submaximal intensity). These and many other factors that are related to the individual's understanding of, or engagement with, a given context may impact on the learning of causal relationships but are rarely captured satisfactorily by formally quantified associative learning mechanisms. These factors all involve prior knowledge of one form or another, and it must be assumed that their influence thus depends very heavily on prior learning. The point of defining them as non-associative is not to make particular claims about their content or the mechanism by which they were initially acquired, but rather to acknowledge that their often substantial impact on new learning is not formally captured in existing models of association formation[1].

BLOCKING IN HUMAN CAUSAL LEARNING AND ITS ASSOCIATIVE EXPLANATION

In the following, we rely heavily on research on the *blocking effect* (Kamin, 1968). In terms of its influence on the development of new theories, Kamin's blocking effect in Pavlovian conditioning is historically one of the most important phenomena. The blocking effect is also an often cited example for a basic learning effect that is regularly reported in human causal learning. It has been one of the empirical cornerstones for the argument that the same underlying learning processes are controlling learning in both conditioning and in human causal learning (Dickinson et al., 1984) and it is therefore not surprising that a particularly high number of studies have investigated non-associative influences on human learning in the context of blocking paradigms.

In a simple blocking experiment (see also **Table 1**), participants might first observe that cue A results in the occurrence of the outcome (A+). In a subsequent phase they may also observe that cues A and B together result in the occurrence of the outcome (AB+). On other trials, they observe that cues C and D together also result in the occurrence of the outcome (CD+). When asked to judge whether B causes the outcome, participants will often give a rating that is substantially lower than their rating for either C or D, even though all three cues have resulted in the outcome on an equal number of occasions. For example, if participants first experience that apples cause an allergic reaction in a Patient X and afterward that two different food combinations, one comprising apples (A) and beans (B) and one carrots (C) and dates (D), will both lead to an allergic reaction, they will rate beans as less likely to cause the allergic reaction on its own than carrots or dates (i.e., B < C or D, see for example, Luque et al., 2013).

TABLE 1 | A typical set of contingencies displayed to participants as part of a blocking experiment.

	Stage 1	Stage 2	Test
Blocking cues	A+	AB+	B?
Control cues		CD+	C? D?

Letters A–D refer to predictive cues, + refers to the presentation of the outcome, ? to the absence of feedback on test.

Figure 1 depicts an example of a simple standard associative network after training and its interaction with the events in the world. The network inside the box on the left side consists of two cues, A and B, which are each connected via associations to the representation of the outcome. As a result, the presence of a predictive cue, represented by the black rectangles outside the box, activates not only its own representation within the associative network but brings also to mind the associatively connected representation of the outcome. Associative models claim that the strength of this associative retrieval of the outcome is a key source of evidence in making a prediction about the outcome. The outcome is rated as likely to occur after certain cues because these cues activate its representation and thus result in an expectation of the outcome. As the strength of this associative retrieval, and thereby the outcome expectation, is a function of the strength of the associative links between the presented cues and the outcome, *f(V)*, differences in responding are based on differences in associative strength.

Most associative accounts furthermore rely on the *prediction error* in some way to establish the associations. Broadly speaking, changes in associative strength, ΔV, are proportional to the error made in the prediction of the outcome, i.e., the violation of the expectation. Every time a prediction is made, it is compared to the actual outcome, represented via the value assigned to λ (e.g., $\lambda = 1$ if the outcome is present; $\lambda = 0$ if absent). The resulting prediction error, that is, the difference between the actual outcome and the outcome expectation, is given by the generalized error term $[\lambda - f(V)]$, and is used to optimize the associative links such that the error is minimized in future predictions. Most models therefore agree that the current prediction plays a key role for the formation and further adjustment of the associative links. Different models assume different ways of combining the associative effects of several cues, that is, when several cues are presented together at the same time, for example A and B in an AB compound. The Rescorla and Wagner (1972) model and most others like it rely on an additivity assumption, that is they assume that the associative effect of cue A and cue B will be the same when they are subsequently presented within the compound AB and will simply sum together, $f(\mathrm{V}) = \Sigma V$. The equation controlling the changes in associative strength can thus be expressed in the following way: $\Delta \mathrm{V} \sim (\lambda - \Sigma \mathrm{V})$.

According to associative theories of blocking, participants will develop a strong association between the cue A (apples) and the outcome (allergic reaction) in the first stage of a blocking experiment, so that, whenever A is presented, the outcome will

[1]There are successful attempts to capture at least some of the effects we refer to as non-associative within a purely associative learning framework. For instance, Haselgrove (2010) has successfully modeled the effects of some forms of pretraining on blocking by appealing to the role played by common elements. However, these attempts rely on assumptions that should not necessarily be easily generalized to other learning designs and paradigms. For instance, a necessary assumption of Haselgrove's (2010) explanation is strong generalization from previously trained cues to the critical test cues and is most clearly applicable to designs in which most or all stimuli are paired with the outcome. In contrast allergist tasks are usually run with multiple filler cues that do not predict the outcome and most participants in these experiments show very clear discrimination between trained and novel cues.

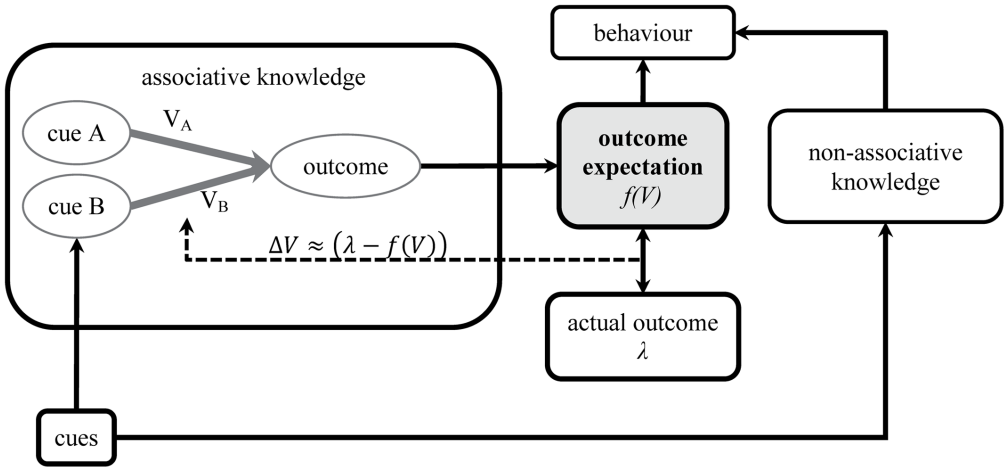


FIGURE 1 | Simple associative network and its interaction with the events in the world.

be retrieved and correctly expected. If the outcome is already expected and subsequently occurs, then minimal learning will take place because the outcome was not surprising (i.e., no prediction error). The same is true for the AB compounds in the second stage. Cue A will again strongly activate the outcome representation and result in a strong expectation of the outcome. Therefore, the prediction error in AB+ trials is already minimal and further correction of the associative links of either A or B with the outcome are soon unnecessary. Learning about B will be blocked by the previous learning about A, and B will not develop a strong association with the outcome. In contrast, the control cues C and D will not retrieve the outcome representation at the beginning of the second stage of the blocking procedure and will therefore not generate a strong expectation of the outcome in CD+ trials. The resulting prediction error will in turn fuel the formation of associative links between C, D and the outcome. At the end of the experiment, C and D will each result in a stronger associative retrieval of the outcome than B (because $V_B < V_C/V_D$) and thus in a stronger prediction, even though they were all paired the same number of times with the outcome. The concept of *prediction error* as a determinant of learning is thus closely linked to the blocking effect and furthermore, prediction error has been shown to be important if not indeed causal for blocking (Steinberg et al., 2013). Blocking is regarded as an instance of cue competition because A and B (and likewise C and D) arguably compete over the association with the outcome. B loses this competition as it is paired with A, which had a head start by virtue of its prior individual pairing with the outcome.

SOME KNOWN EFFECTS OF NON-ASSOCIATIVE KNOWLEDGE

From the description of a simple associative network, it should be apparent that participants that receive the same cue-outcome pairings should show the same learning as this is the only kind of information on which the formation of associations and thus the expectation is based. As already pointed out, however, it is well established that non-associative knowledge affects learning and decision making. One classic demonstration of the effect of verbal information on causal learning was provided by Waldmann and Holyoak (1992). Their experiments were designed to create two learning tasks that were equivalent at the associative level – that is, identical in terms of the statistics of the events involved – but differed in terms of the general causal information conveyed in the cover story. All participants in Waldmann and Holyoak's study received the same cue-outcome combinations during Stages 1 and 2 of a blocking experiment. However, the cover story established either a predictive or diagnostic learning situation for these cue-outcome pairings. While the cues were always the same, participants in their predictive task had to learn which cues would *cause* a new kind of emotional response in observers. In contrast, participants in the diagnostic task saw the same cues but redefined as symptoms of a disease and had to learn which symptoms *were caused* by the disease. Even though subjects saw identical cues and cue-outcome pairings, they rated the critical target cue differently in the diagnostic and in the predictive condition. Specifically, participants given the diagnostic scenario gave the target cue, B, a stronger rating. As Waldman and Holyoak's study did not include the appropriate control cues for blocking, C and D, drawing a conclusion on the blocking effect is not possible. However, similar subsequent experiments have replicated the effect of causal model, implemented through instructions and prior knowledge, on blocking and other effects (e.g., Waldmann, 2000, 2001; Luque et al., 2008; Blanco et al., 2014; but see Shanks and Lopez, 1996; Thorwart and Lachnit, 2010).

Another line of experiments has demonstrated an effect of inferential reasoning on blocking. These experiments show how information about the causal relationship between cues and outcomes influences learning. In De Houwer et al. (2002) experiments, participants had to rate how likely it was that a tank would be destroyed (i.e., the outcome) if a certain weapon was fired (i.e., a causal cue) or if an indicator lit up (i.e., a

predictive cue). Weapons and indicators were represented by the same abstract visual cues that were present during training shortly before the possible destruction. Nevertheless, participants rated the relationship between them and the destruction of a tank differently. Beckers et al. (2005b) replicated this result in 4-year-old and 8-year-old children with a scenario about predicting rain, rather than exploding tanks. A related series of studies addressed how assumptions about the additivity of the causal effects of cues may determine the strength of the blocking effect (Lovibond et al., 2003; Beckers et al., 2005a). An additivity rule would state that, if two cues cause the outcome separately, the outcome should be even stronger when both cues are present at the same time. This assumption permits the application of simple deduction such that on observing that A and B together do not cause a more severe allergy than A alone, B must therefore not cause the outcome. Consequently one would expect to observe blocking from inferential reasoning alone if participants hold this assumption. A non-additivity rule would instead reflect the belief that adding a second cue does not increase the likelihood or strength of the outcome if this is already predicted by the first cue, even if both cues are predictive on their own. Pretraining and verbal instructions were successfully employed to shift participants' beliefs in one or the other direction and these experiments showed repeatedly that affirming the additivity rule strengthens the blocking effect.

Models of associative learning are designed to account for blocking but, as these examples show, the presence of the effect itself varies considerably across procedures, and in ways that seem to be more consistent with cognitive processes that differ considerably from the simple principle that learning is proportional to prediction error. For the theoretical approach typified by associative networks, the challenge posed by these results is not the fact that they show other cognitive factors play a role in controlling behavior. For instance, nowhere have associative models explicitly assumed that other mental processes cannot produce cue competition effects (symbolized by the arrow from cues to non-associative knowledge in **Figure 1**). But the fact that associative models do not speak to these non-associative factors works against their relative utility as accounts of human learning, since there are clearly important properties of human learning, judgment, and behavior that they fail to capture. Clear evidence exists that non-associative factors influence associative learning in the laboratory. It *might* be the case that this evidence reflects a thin veneer of cognitive penetrability on an otherwise highly regular and lawful set of learning principles that capture real-life learning quite well. After all, knowledge that one is participating in a psychology experiment must surely encourage introspection and careful thought. Alternatively, and more worryingly for the conventional associative approach, this evidence may be symptomatic of broad, general and far-reaching sensitivity to a host of factors that are poorly accounted for by associative learning networks. Therefore, even if one is to retain the association-formation approach in theorizing about human learning, there is a need to better understand how other factors play a role in human learning.

HOW MIGHT NON-ASSOCIATIVE KNOWLEDGE INFLUENCE AN ASSOCIATIVE NETWORK?

Since non-associative knowledge can clearly influence associative learning phenomena, including those that form the basis of contemporary prediction-error models, it is tempting to discard the notion that we possess a system dedicated to mental association in the manner described by associative networks. Indeed some authors have already reached this conclusion (Mitchell et al., 2009). They too assume that the expectation of the outcome will inform our behavior, but this expectation is based on generating and evaluating propositions in deductive reasoning processes. However, an alternative approach, and one that we think is still instructive, is to ask how non-associative knowledge could impact upon learning, expectations, and behavior if we assumed that a general-purpose associative system was still in place. How could the non-associative knowledge influence the operations of such a system and what would be the implications if it did so? Here we consider briefly several possible ways in which this could occur.

Non-associative Knowledge May Change the Inputs to an Associative Network

One might account for the variations in the blocking effect in causal learning by suggesting that the formation of associations is sensitive to parametric differences. Certainly, most associative learning models generate parameter-specific predictions about various learning phenomena, meaning that quantitative parameter variations can produce different effects without fundamentally changing the inner workings of the network itself. They affect what is happening but not how it is happening. Often the manner in which stimuli are represented within an associative network can be critical for how learning takes place. For instance, although associative learning models generally predict blocking, the strength of the predicted effect can vary widely according to assumptions about how the stimuli are mentally represented and how quickly learning occurs during training. These assumptions are captured in parameters like the associability of the cues. If non-associative knowledge alters such parameter values, it would affect learning without replacing or even fundamentally changing the learning mechanism that is assumed to have worked and survived successfully throughout evolution. However, why and how should non-associative knowledge influence quantitative parameters of the network?

Associative models generally assume that physical properties of the cues and the outcome influence parameters like their associability such that there is a link between basic perceptual principles and the determinants of learning (Annau and Kamin, 1961; Mackintosh, 1976; Redhead and Pearce, 1995). Many theorists take a further logical step by assuming that basic cognitive operations like attention also determine key aspects of stimulus representation in the learning system. That is, the mental representations that engage in learning reflect information subjected to limited sensory processing, which is selectively biased by attention. Theorists have often assumed that

selective attention affects learning in other animals just as it appears to in humans (e.g., Lashley and Wade, 1946; Sutherland and Mackintosh, 1971; Mackintosh, 1975; Pearce and Hall, 1980). Therefore, to the extent that beliefs derived from non-associative knowledge can affect attention and perception, those beliefs may also impact upon learning within an associative network, even if the operations of that network are relatively automatic (**Figure 2**).

Work on the learned predictiveness effect clearly demonstrates an effect of instructed attention on selective learning (Mitchell et al., 2012; Don and Livesey, 2015; Shone et al., 2015). The learned predictiveness effect is a widely observed learning bias toward previously predictive cues in novel situations (see Le Pelley et al., 2016 for a recent review). The effect is generally attributed to an attentional shift that occurs as a natural consequence of acquiring associative knowledge about those cues. As cues are associated with outcomes, attention to more predictive cues is enhanced, resulting in faster learning for those cues within the associative network. However, Mitchell et al. (2012) demonstrated that explicit instructions manipulating participants' beliefs about the predictiveness of cues in a second learning phase had significant effects on their learning of cue–outcome contingencies. After learning about the predictiveness of the stimuli in a trial-and-error fashion (which we assume led to acquisition of associative knowledge), participants in a "continuity" group received instructions at the start of Phase 2 that the cues that were predictive in Phase 1 were also likely to be predictive in Phase 2. A "change" group received opposing instructions, that the cues that were predictive in Phase 1 were unlikely to be predictive in Phase 2. Interestingly, participants in the "change" group showed a strong reversal of the learned predictiveness effect. That is, more was learned about previously non-predictive cues than previously predictive cues in Phase 2. Subsequent studies have partially replicated this sensitivity to instructions, though have typically found much weaker instructed reversal effects accompanied by a continued influence of biases established in Phase 1, despite clear evidence that the participants have read and understood the instructions (Don and Livesey, 2015; Shone et al., 2015). While Mitchell et al. (2012) favored an explanation purely based on conscious reasoning processes, where participants deliberately attend to the cues they believe are important, a viable alternative is that attentional processes are brought under conscious control and thus let non-associative knowledge influence the course of subsequent learning. This source of influence does not necessitate that non-associative expectations fundamentally change the operations of the associative network itself, merely what it receives (Livesey and Harris, 2009). In other words, a cue that possesses relevance merely because the instructions have enhanced its importance may be better or more fully represented in an associative network (i.e., have greater salience) because the individual is deliberately attending to it.

This might also go some way to explain some instances where the blocking effect appears to be unreliable or completely absent. In addition to the associative processes explained above, some theories assume that blocking is partly governed by a lack of selective attention to the blocked cue, either because it is redundant (Mackintosh, 1975) or because the outcome is predictable (Pearce and Hall, 1980). If non-associative factors influence selective attention, they may provide a means by which attention to the blocked cue is enhanced (or reduced even further), which could alter the likelihood of observing a blocking effect considerably even if learning were still primarily based on association formation.

In addition, if non-associative knowledge can affect the way stimuli are represented then this knowledge may also change the manner in which associative retrieval generalizes from A to AB at the beginning of Stage 2 learning and from the compounds to the single stimuli presented on test (Livesey and Boakes, 2004; Thorwart and Lachnit, 2009, 2010). Several authors have suggested that pretraining, task instructions, and spatial stimulus characteristics can alter the encoding strategy that participants use or the way they mentally represent cues, which in turn affects

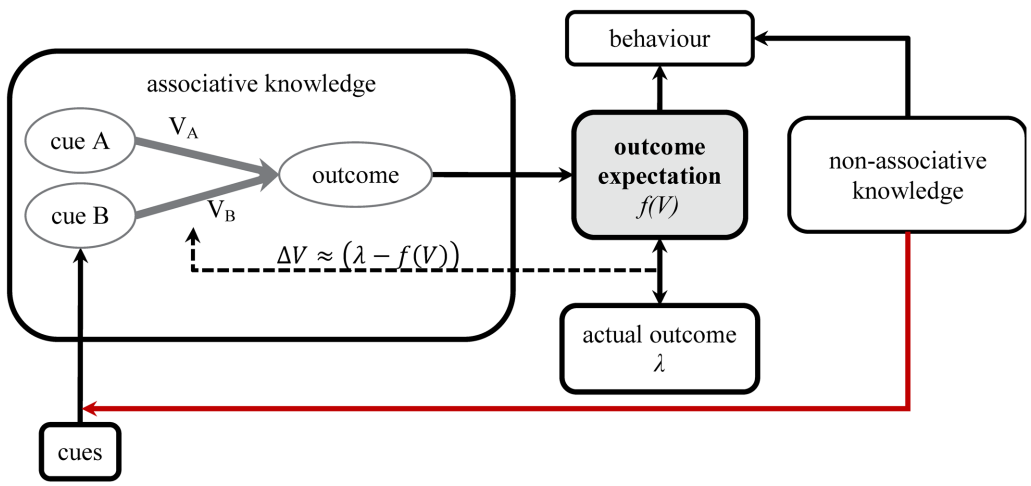


FIGURE 2 | Non-associative knowledge may change the inputs to an associative network (indicated by connection in red).

generalization between compounds and individual cues (e.g., see Melchers et al., 2008 for a review). The potential for these changes in stimulus representation to impact on learning is sometimes discussed in terms of flexible shifting between elemental and configural learning (Melchers et al., 2008) or shifts within an elemental learning system (e.g., Wagner and Brandon, 2001; Livesey and Harris, 2008; Thorwart et al., 2012, 2016). Such changes in stimulus representation reduce generalization from A to AB and thus result in a weak expectation of the outcome in AB+ trials. The resulting increased prediction error supports considerable further correction of the associative links of both A and B with the outcome. This change in stimulus encoding would also affect the generalization from trained compounds AB+ and CD+ to individual test cues B, C, and D, which may result in overall weak expectation and a smaller blocking effect in test, where blocking is generally measured by the difference between the rating of B and the mean rating of C and D. If all ratings are low due to reduced generalization from the training compounds AB and CD, the blocking effect will be small as well. No matter how the global properties of stimulus representation operate, the broader issue at hand is that generalization between different trial types might vary according to various sources of non-associative knowledge that affect stimulus encoding, which in turn impact on the expectation of the outcome when a new but related trial type (e.g., AB+) is experienced.

Finally, how and what information is sampled by the learner affects learning (Matute, 1996; see Fiedler and Juslin, 2006, for similar arguments in relation to decision making) and it is known that sampling strategies can be modified through verbal instructions (Matute, 1996; Blanco et al., 2012) or the amount of personal involvement (Yarritu et al., 2014). This influence is clearest in instrumental tasks where the learner's actions directly control the delivery of outcomes and thus also the opportunities to observe relationships between action and outcome. For instance, in contingency judgment experiments where participants are asked to judge the degree of control of an action over the occurrence of an outcome, participants often perform the action relatively frequently (e.g., on considerably more than 50% of trials), which in turn limits the opportunity to learn about the likelihood of the outcome in the absence of the action and creates circumstances that favor overestimation of the association between the action and outcome. Changes in action strategy can thus directly influence the quality of the evidence for statistical relationships between events, and these strategic changes could be initiated by any number of non-associative manipulations.

Non-associative Knowledge May Change How Associative Outputs Translate to Beliefs and Behavior

The clearest evidence that associative and non-associative knowledge might provide dissociable expectations at a behavioral level comes from studies that compare explicit predictions and ratings with other behavioral measures such as response priming and conditioned responding that gauge expectation less directly. One example is the Perruchet (1985, 2015) effect, where within the same experiment and indeed the same trial, diverging response patterns can be obtained in two behavioral systems (for example eye blink conditioning and causal rating; for details see below). Cobos et al. (2016) observed diverging "associative" and "non-associative" response generalization in cued response times and verbal ratings, respectively and Morís et al. (2014, Exp 4) found that non-associative knowledge, given by instruction, affects verbal judgements but not responses in a recognition priming-based test. But a related and in many ways more difficult question is how associative predictions might generate explicit judgements.

Most associative models generate predictions about behavior based on the summed associative strength of the cues that are active, or the activation of the representation of the outcome itself, outputs that we will refer to as *associative predictions*. Because they are usually intended to apply to a wide range of behavioral paradigms, few associative models provide formal rules for translating these associative predictions into specific behaviors. Fewer still provide precise rules for how associative predictions should be translated into judgements or verbal behaviors of the variety that can only be meaningfully measured in human learning. As such, when model predictions are tested empirically, they are usually expressed as ordinal hypotheses rather than precisely quantified predictions.

A problem thus still remains in characterizing how associative predictions are conveyed in the explicit expectations of the individual and whether the relationship between the two should be expected to be consistent across different experimental situations. It might well be expected that simple memory and retrieval mechanisms determine our judgements in at least some situations. The classical associationist view is that a cue might be judged as being the cause of an outcome to the extent that the presence of the cue brings to mind the idea of the outcome. Similarly we might expect a particular outcome to occur simply because a representation of that outcome has been activated via its associations with other cues that are present at the time.

Theoretically, this relationship between associative retrieval and causal rating could be regarded as an immutable property of a system that integrates memory with an understanding of causal structure. Alternatively, it may be that the fluency of memory retrieval serves as just one source of evidence on which judgements about causation and expectations about future events are based, as conclusions based on non-associative knowledge serves as a second (**Figure 3**). In some circumstances, associative activation of the outcome may form the strongest available evidence about what is going to happen when a cue is presented, or the strongest indicator of how the individual should behave. But under other circumstances, for instance where it is very clear that a deductive reasoning process should be used, associative memory retrieval may play a relatively minor role. Thus the relative strength of non-associative knowledge may play an important role in how associative predictions translate to overt judgements and predictions. One might then assume that associative learning in the form characterized by associative networks exists and operates fairly consistently across different individuals and contexts, and that most of the variance in causal judgments results from non-associative factors having

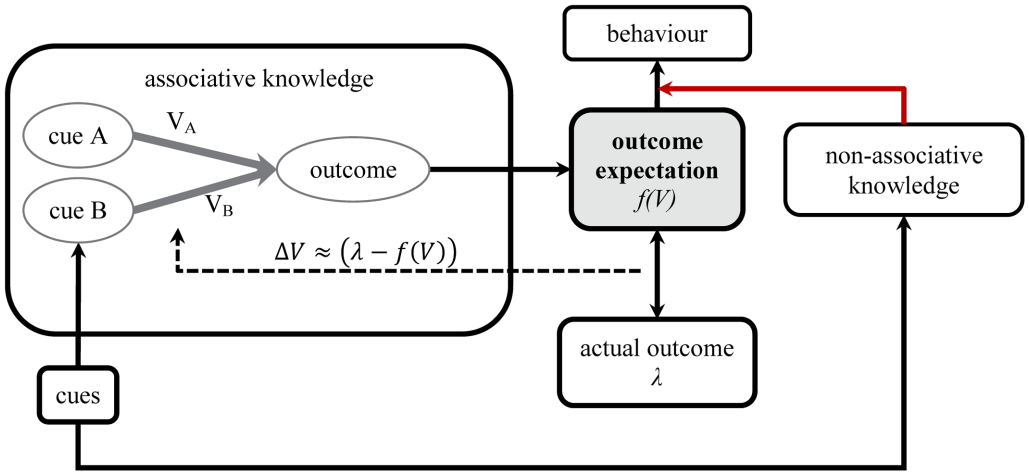


FIGURE 3 | Non-associative knowledge may change how associative outputs translate to beliefs and behavior (indicated by connection in red).

an influence on performance, for instance in the interpretation of associative memories and their translation into explicit behavior.

This possibility again does not imply that the internal workings of the associative network need be materially affected by expectations derived from non-associative knowledge. It merely assumes that associative predictions do not always have a strong influence on behavior. Returning to the blocking example, it is possible that the observed sensitivity of predictive ratings to non-associative information about causality (e.g., blocking is more readily observed in causal scenarios than non-causal scenarios) means that associative retrieval plays no part in determining the judgements made in either scenario. But it could also mean that associative retrieval plays a greater a role under some instructional and task conditions than others (e.g., Sternberg and McClelland, 2012). For instance, perhaps judgments that feel more naturally intuitive or familiar to the individual allow a greater influence of associative predictions, particularly among individuals who are disposed to making intuitive judgments already (Livesey et al., 2013). Support for such an influence of non-associative knowledge may be found in studies by Matute et al. (1996), Vadillo et al. (2005) and Vadillo and Matute (2007), which showed that the precise wording of the test question does have an influence on judgements. For example, Matute et al. (1996) found that the relative-validity effect, another cue competition effect related to blocking, appears when subjects are asked to rate whether the target cue X is a cause or an indicator of the outcome, but vanishes when participants are asked to rate to what extent cue X and the outcome co-occurred. Similarly, Gredebäck et al. (2000) found a significant cue competition effect when participants were asked about the predictive value of the cue, as well as when they were asked about the causal relationship between the cue and the outcome. However, the cue competition effect did not reach statistical significance when participants were asked about the probability of the outcome given the cue, nor when they were asked about the frequency of cue–outcome pairings.

Many of the results that we have discussed thus far, including those that show a sensitivity of blocking to causal model, contain single dissociations in which the behavioral ratings in one condition are generally closer to ceiling (e.g., Waldmann, 2001) and therefore change the likelihood of observing differences between ratings for reasons that might be to do with the measurement scale rather than the underlying process. For example, ratings in non-causal scenarios tend not to show blocking effects as readily as causal scenarios, specifically because the rating for the blocked cue is higher. If there were differences in how participants regard the blocked cue and the control cues that were in fact equivalent under causal and non-causal scenarios, it is reasonable to assume that those differences would appear to be weaker, possibly even non-existent, if ratings were generally near ceiling anyway. Thus an observation that blocking is weaker in non-causal scenarios could be achieved simply by assuming that participants use the scale differently in the two scenarios, without making any assumptions about changes in underlying process. Although we do not necessarily favor an explanation purely in these terms, it is worth pointing out that the evidence suggesting sensitivity to non-associative influences on causal learning is often consistent with multiple explanations, and at least some of these explanations do not assume that anything fundamentally different is happening in terms of learning and memory when non-associative knowledge is manipulated.

Non-associative Knowledge May Influence Association Formation Directly

Assuming that associative learning *does* occur via an associative network of some form, the previous two hypotheses do not necessitate that non-associative cognitive processes have any direct impact on how associations form within that network. Rather, they may affect the information that is fed in to the network and what is done with the output that the network returns. One could posit that cognitive interactions of these forms occur and still assume that associative learning is relatively

modular in its operations. However, it is worth considering an alternative hypothesis in which learning within associative networks is directly affected by non-associative factors.

Outcome expectation and prediction error form the centerpiece of many associative learning rules and the obvious and most effective point of interaction of non-associative knowledge with associative processes. Changes to the outcome expectation have profound effects on the updating and thus the structure of the associative network representing the relationships between current cues and outcomes, even if the outcome expectation is more a result of the associative learning system than a part of it.

Since associative learning is often assumed to be proportional to prediction error and predictions can often be made on the basis of both associative and non-associative information, an obvious way in which a direct non-associative influence might occur would be if prediction error was a function of all sources of outcome expectation, and not just associative prediction. In this case, controlled cognitive operations based in metacognition and reasoning could have a significant impact on a key variable that determines trial-to-trial variations in associative learning. Thus variations in cognitive processes could have a lasting impact on the course of associative learning even though association formation lawfully follows a basic learning rule.

Figure 4 shows how this interaction of non-associative knowledge with the associative network could work. Crucially, the associative network does not contain additional or enriched representations of information about the cover-story or the outcome and the links are still simple quantitative links that contain no qualitative or structural information about the relationship between the events. It is also important to note that there are no higher-order deductive reasoning processes assumed to be responsible for optimizing and changing the network. Indeed, the associative network in the box in **Figure 4** is exactly the same as that in **Figure 1**, and when a cue is presented, it activates its associative representation and all associatively linked events. The activation of the outcome representation therefore depends on the strength of the associative link and we assume that its retrieval remains a key source of evidence in deciding whether to predict the outcome or not.

However, it is not the only source of evidence as the cues, the learning situation, or the retrieval of the outcome representation itself can trigger other mental processes. After the associative retrieval of the outcome, this knowledge is used to re-evaluate and adjust the expectation of the outcome. The final outcome expectation is then a function of both the strength of the associative links between the presented cues and the outcome and any other information that the learner perceives as being relevant [*f(V, other)*].

One source of non-associative influence is the extent to which the individual reflects upon their own learning and thought processes, that is metacognitive processes. This may be a strong source of variance across different procedures and across individuals and if associative learning is sensitive to the operations of metacognition (in any one of the ways outlined earlier) then this could be a major source of variance in cue competition and other learning phenomena. An obvious way in which metacognition may be relevant to prediction error is the possibility that associative predictions are evaluated and potentially revised by the individual prior to observing the relevant outcome. We describe this re-evaluation as being metacognitive as it relies on assessment of the outcome expectation and some cognizance of the source of that expectation. Thus we typify the process as being very explicit and probably quite variable between individuals and between learning contexts. We will consider an example in relation to blocking.

Associative learning models all assume some degree of generalization between trials that have cues in common. In the case of blocking, pretraining with A+ leads to an expectation of the outcome in the presence of A. This expectation generalizes to AB+ trials. As described above, the default assumption of many associative learning theories is that the associative strengths of the cues that are present will combine in an additive fashion (Rescorla and Wagner, 1972), although there are many hypothesized reasons why this summation might be less than perfectly additive (see McLaren and Mackintosh, 2000, 2002; Wagner and Brandon, 2001; Harris, 2006; Harris and Livesey, 2010; Thorwart et al., 2012 to name just a few). Thus the process that provides a means of generalization is assumed to automatically produce an expectation of the outcome based on some combination of the associative strengths of the cues present. This assumption is based partly on direct evidence of summation in human and animal learning (see for example Myers et al., 2001; Pearce, 2002; Soto et al., 2009; Thorwart et al., 2016) but also on the fact that it is necessary for the associative account of the blocking effect and that the blocking effect is found in diverse and various circumstances and paradigms, indicating that the additivity rule is in fact the default mode by which our learning system operates.

In contrast, when an individual is deliberately engaged in the task of trying to understand the general rules by which relationships between cues and outcomes abide, they may have reason to question this simple summative principle and they may do so to differing to degrees depending on the individual and the context in which they experience the cues and outcomes. We might assume that the process operates according to the following. In phase 1 of a blocking experiment, in addition to forming an association between A and the outcome, the participant has episodic memory of witnessing certain trial types (e.g., A+) and entertains beliefs about the relationship between A and the outcome. In the second stage, the current trial type (AB+) has some overlap with previous experience and associative memory results in retrieval of the outcome representation or increased activation of the outcome representation. This would normally result in an associatively retrieved expectation that the outcome will occur. The learner might accept this expectation at face value and thus will not be particularly surprised to find that the outcome occurs again on this new AB trial. However, the participant may also notice that the current trial type (AB+) is not the same as those previously witnessed. Although the participant has strongly retrieved the outcome representation, they might question whether their expectation of its occurrence is accurate given their uncertainty about the perceived change in trial type. The learner may acknowledge the fact that this is a novel situation, that they don't know

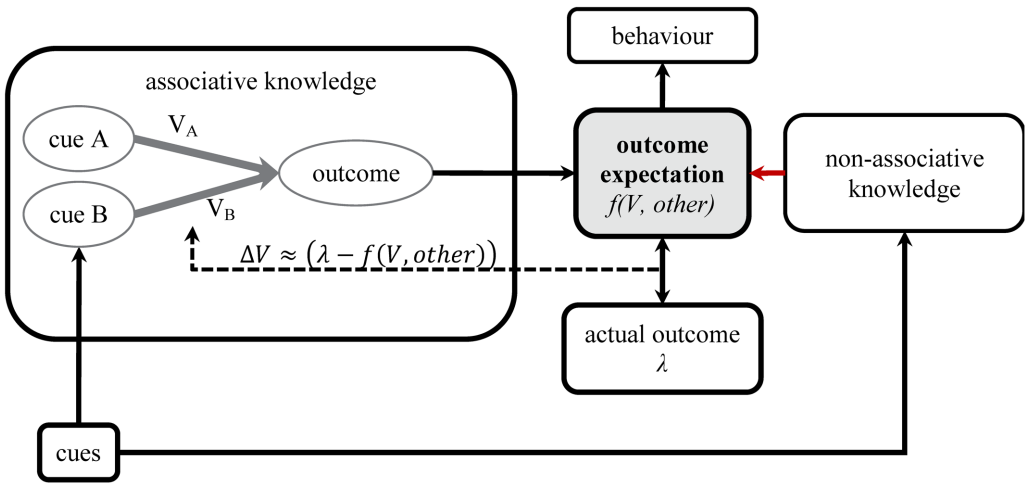


FIGURE 4 | Non-associative knowledge may influence the outcome expectation that is directly involved in association formation.

how these indicators operate in combination and entertain the possibility that indicators A and B together might not indicate the same outcome as A alone. The expectation generated on the basis of A+ episodes is consequently moderated, and the learner may regulate their predictions in a way that reduces their expectation of the outcome. That reduction affects both the explicit predictions of the individual and the associative learning that takes place when the participant observes the outcome on that trial. This cautious approach means that the occurrence of the outcome on such trials is still at least partially surprising and its presence should be learned about more effectively. Thus, an associative link between B and the outcome may be established and the blocking effect attenuated.

At face value, this is simply a cognitive description of *external inhibition*, a well-documented effect in animal learning (e.g., Pavlov, 1927) in which the addition of a novel cue reduces the learned response to a previously trained cue. The difference here is that we specifically assume that moderation occurs as a consequence of the participant's appraisal of what they know about how the cues and outcomes generally operate. Consequently, one can begin to predict different effects on learning in conditions where the cues and outcomes are the same but the causal scenario differs. In a food allergist experiment, a participant might first observe on multiple occasions that their "patient" has consumed Apple and suffers an allergic reaction as a consequence. From this they may form an association between Apple and the reaction, and they may also form a belief that the patient is allergic to Apple. When the patient then eats Apple and Beans in one meal, it seems reasonable to assume that most participants would believe that the patient will suffer an allergic reaction because Apple was eaten. But what of a situation in which the cues are unknown drugs that cause or prevent side effects, or symptoms of a hitherto unknown disease? Is it reasonable to assume that if a patient suffers a migraine after being given Melixil, they will also suffer a migraine when given Melixil and Andrum? Many people might be considerably less sure of this, given that they know nothing about the drugs and have little relevant experience to draw on. One might therefore assume that the expectation of the outcome generated by Melixil (cue A) will be moderated by the uncertainty that the individual feels about the scenario, about the way cues interact, or the reliability of their effects (using this drug scenario, Lee and Livesey, 2012 found no evidence of blocking).

The hypothesis being entertained here is that uncertainty about new trial types may increase the amount that is learned about a redundant cue. The assumption is that factors that increase the uncertainty of a participant about the current learning situation decreases blocking. An extension of this hypothesis would further predict that participants will learn less about the blocking cue B when their natural assumptions about cause and effect in a given scenario are not contradicted by instructions or pretraining. That is, if the participant *feels* well-informed and confident about their understanding of the situation, they may show less evidence of learning about redundant cues.

A hypothesis of this sort is applicable to the influence of non-associative knowledge about the additivity of outcome properties, and specifically how this impacts cue competition. Most causal judgment experiments present deterministic relationships where the probability of the outcome is either 0 or 1 depending on the cue or cues presented, and the presentation of the outcome consists of little more than a label or picture. Therefore the method of presenting the outcome to participants lacks the clarity of information needed to determine whether the outcome is truly additive or non-additive. Lovibond et al. (2003) suggested that this is the reason that blocking is typically fairly weak in human causal learning experiments, because not all participants maintain an assumption of outcome additivity during the experiment. They set about testing the effect of outcome additivity assumptions by giving one group of participants pretraining that explicitly demonstrated the additive nature of the outcome and another group of participants explicit pretraining

demonstrating that the outcome was the same magnitude whether there were one or two causes present. The additive group received pretraining in which two cues, which were unconnected to the cues A and B of the actual blocking training, each led to the outcome (X+Y+) and their compound led to an even stronger outcome (XY++, e.g., a stronger allergic response). This group subsequently displayed significantly larger blocking than the non-additive group, which received pre-training in which the compound led to the same outcome as the single cues (X+ Y+ XY+). This result has been replicated in several studies (e.g., Livesey and Boakes, 2004; Beckers et al., 2005a; Mitchell et al., 2005).

This is problematic for associative accounts as no associative knowledge about A and B is established in the pretraining. A common explanation offered is that additivity assumptions encourage deductive reasoning, which results in a conclusion that the blocked cue is not a cause of the outcome (e.g., see Mitchell and Lovibond, 2002; Lovibond et al., 2003; Beckers et al., 2005a). While this explanation is certainly very plausible, additive pretraining like X+ Y+ XY++, which is usually accompanied by very explicit instructions about cue additivity, also removes any uncertainty about the way cues combine in a particular learning situation and in this way could influence the outcome prediction and therefore the prediction error on AB+ trials. As A is known to lead to the outcome, the learner will indeed be unsurprised to find that the outcome occurs again on this new AB trial and no prediction error will occur. However, after non-additive X+Y+ XY+ training, participants still know very little about the way the cues combine. The participant may entertain the hypothesis that the influence of the cues is somehow normalized or that there is a ceiling effect masking the summative effects. If uncertainty at a metacognitive level reduces the outcome expectation, prediction error will increase when AB+ trials are experienced and thus more associative learning takes place when the participant observes the outcome on that trial.

One result that clearly conflicts with this explanation is Beckers et al.'s (2005a) finding that manipulating assumptions about additivity *after* the trial-by-trial learning has already taken place still influences the strength of the blocking effect. It is clearly implausible that the operations of an associative network at the time of learning could be influenced by this later non-associative knowledge. However, non-associative knowledge does not need to change the operations of an associative network at the time of learning but only the impact of the associative knowledge on performance in the test phase after learning, either by influencing the outcome expectation directly or by changing the expression of the associative prediction, as described above. The experience of additional cues between training and test might increase the influence of non-associative knowledge on the outcome expectation by increasing uncertainty – if only for the additional time that has passed between training of the compounds AB+ and CD+ and testing the cues B, C, and D. In this case, blocking under the additive condition may be enhanced because causal ratings for the cues are only weakly related to associative memory and are moderated by the reasoning that additivity instructions strongly encourage.

We have described how an unfamiliar context or unfamiliar cues like unknown drug names will increase the uncertainty of learning situation and how this can explain why it is much harder to find blocking in one scenario than in another. In Waldmann and Holyoak (1992), participants showed less blocking in the diagnostic than in the predictive condition. While the cues were always the same stimuli, participants in their predictive task had to learn whether certain cues would elicit a new kind of emotional response in observers. In contrast, participants in the diagnostic task saw the same features redefined as symptoms of a disease and had to learn which symptoms were diagnostic for the disease. We would argue that the diagnostic learning situation increased uncertainty, for instance because the cover story established that the outcome actually precedes the cues in real life, so that participants were in a situation where they had to "predict" an outcome that had already happened. Furthermore, participants in the diagnostic situation have to take into account alternative diseases as causes of the observed symptoms (Waldmann, 2001). For example, even though fever may be an effect of flu, it has many alternative causes, which participants cannot rule out easily within the learning situation and thus increase the uncertainty about their prediction.

ISSUES, LIMITATIONS, AND FUTURE DIRECTIONS

The scope of our discussion has been necessarily highly selective and has avoided several issues that are obviously important. As we have noted, we make no attempt here to specify in any way how non-associative knowledge is acquired, and define it simply as cognitive influences that associative networks make no attempt to explain. This undoubtedly belies the complexities involved in acquiring such information. In describing three basic ways how non-associative knowledge might influence learning in an associative learning system, we have also avoided consideration of how their effects might combine. It might well be that sources of non-associative knowledge influence the processing of the cues, the translation of the outcome expectation in behavior as well as the expectation of the outcome directly at the same time. However, for sake of the theoretical exercise, we have left the interaction of all three possible mechanisms out of consideration.

We have chosen to focus our discussion on results from causal and contingency learning paradigms. These results, among others, established the relevance of non-associative knowledge in human causal learning. We would argue that the setting of contingency and causal experiments makes them particularly receptive to such information because they typically rely on explicit and self-paced judgements and since they usually invite the individual to entertain a fictitious scenario in which their previous knowledge may come to bear (even though participants are usually encouraged to ignore what they know about similar causal relationships in the real world). In classical conditioning studies, the experimental situation does not contain much non-associative information that could show an influence on learning. In the extreme case, participants are given no other instruction than to sit in front of a computer screen and pay close attention

to it. Far more contextual information is given in human causal learning studies and the experimental situation is thus more likely to encourage activation of non-associative knowledge. However, this does not mean that beliefs and expectations based on non-associative knowledge do not affect classical conditioning and other forms of human learning. At least some studies support the notion that non-associative knowledge affects the learning of conditioned responses as well. For instance, Mitchell and Lovibond (2002) showed that skin conductance conditioning is sensitive to information about outcome additivity given in the verbal instructions. They observed significant effects only when participants received verbal instructions emphasizing the additivity rule whereas blocking was not evident when the instruction introduced a non-additivity rule. Therefore, we assume that these issues are relevant to all forms of human associative learning and extend beyond the limited selection of procedures and phenomena that we have discussed here.

The account we offer here necessarily involves non-associative processes impacting upon observable behavior (i.e., performance) as well as on the formation of associations (i.e., learning). As such, the large number of studies exploring non-associative factors in associative learning – many of which show that instructions, pretraining, and cover stories affect causal and contingency learning – do not offer unique support for, or refutation of, this approach because most can be explained in terms of a performance-level effect alone. To properly test the hypotheses outlined above, a different approach is required, one in which performance-level and learning-level influences can be dissociated. Applying this logic to blocking in causal learning, instructional manipulations are required which can be expected to change participants' predictions during learning of the AB+ compound *without* resulting in global changes to the way the ratings scale is used at test. There is also still a general need to examine how potential differences in learning manifest differently depending on the properties of the test measure. Although recent work has revealed much about the way blocking is sensitive to causal assumptions, researchers have typically been less concerned with the general properties of the measure itself, even though these properties may strongly affect the potential to observe cue competition effects. The presence of ceiling effects on the strength of ratings provides a simple example of this. As previously noted, using a test measure in which ratings are generally close to ceiling could mask a blocking effect in non-causal scenarios even if the causal scenario made no difference to the strength of learning about competing cues. This simple possibility alone is cause to think seriously about the basic properties of the test measure and is indicative of a more general problem with comparing blocking effects across different conditions. After all, the magnitude of blocking is a *difference* between the judgments made for two types of cue (blocked vs. control), and is often measured on a ratings scale with unknown psychometric properties. Comparing the magnitude of two differences on a measurement scale that is at best ordinal in nature is a risky exercise.

Beyond cue competition, procedures in which associative predictions and non-associative expectation can be directly pitted against each other may be particularly useful for testing the hypotheses outlined in this article. As mentioned above, such examples do exist, though they are relatively rare. Two that might prove useful are Perruchet's (1985) dissociation between the strength of anticipatory responding and explicit ratings of outcome expectancy and Shanks and Darby's (1998) dissociation between similarity-based and rule-based generalization.

Perruchet's dissociation emerged originally in classical human eye-blink conditioning. Perruchet (1985) arranged a partial reinforcement schedule in which the same tone cue played on every trial, but was followed on just 50% of trials by the outcome – an irritant (a puff of air delivered to the eye) that elicits an eyeblink. A conditioning procedure of this kind usually leads to the development of anticipatory eyeblinks during the tone cue in expectation of the airpuff. The randomization of the two trial types (cue-outcome and cue-alone) meant that the trial types sometimes remained the same over several consecutive trials, and sometimes alternated frequently, resulting in short runs of just one or two of the same trial type. When Perruchet arranged the analysis based on the length of the preceding run of trials, he found a pattern of anticipatory eyeblinks that followed the pattern one would expect from conditioning based on basic associative principles. Runs of cue-outcome trials increased anticipatory behavior as a function of the length of the run, whereas runs of cue-alone trials decreased anticipatory behavior as a function of the run length. However, when he asked participants to indicate explicitly how much they expected the airpuff on the next trial, their pattern of expectancies was the opposite; Runs of cue-outcome trials decreased expectancy ratings as a function of run length, whereas runs of cue-alone trials increased expectancy ratings as a function of the run length. This pattern follows a classic gambler's fallacy effect and is inconsistent with the predictions of associative networks. The result has now been replicated across several paradigms involving classical conditioning and voluntary responding (see Perruchet, 2015 for a review). Current debates about the validity of this dissociation center around whether the pattern observed in anticipatory behavior is a *bona fide* example of associative learning (e.g., Weidemann et al., 2009, 2016; Barrett and Livesey, 2010; Mitchell et al., 2010) and whether participants truly hold these two conflicting belief biases concurrently (Livesey and Costa, 2014; Lee Cheong Lem et al., 2015). However, to date there has been no attempt to explore how these beliefs affect future learning. For instance, after a long run of trials on which the outcome has occurred, if another cue-outcome pairing occurs then the prediction error based on associative mechanisms should be relatively small but prediction error based on explicit expectancy should be relatively high.

The Shanks-Darby patterning task was developed specifically to create opposing influences on generalization within a causal learning task. Shanks and Darby (1998) trained participants to solve multiple examples of a positive patterning (e.g., A−/B−/AB+) and negative patterning (e.g., C+/D+/CD−) in a simple food allergist causal learning procedure. In animal learning, conditional discriminations of this variety, and particularly negative patterning, are relatively difficult to acquire (e.g., Harris et al., 2008), and there is at least some evidence that humans too find negative patterning more difficult to learn

than positive patterning (Livesey et al., 2011; Thorwart et al., 2016). Associative networks generally anticipate this difference because the summation of associations formed to the single stimuli in negative patterning (C+ and D+ trials) provides a particularly strong and incorrect prediction for the compound (CD−). However, from an abstract relational perspective, positive and negative patterning possess the same complexity; they are perfect examples for a simple rule that the outcome of the compound is always the opposite of the outcome of the single cues (Shanks and Darby, 1998; Lachnit et al., 2001, 2002; Harris and Livesey, 2008; Cobos et al., 2016). Capitalizing on this simple relational property, Shanks and Darby also trained participants on a series of single cues (I+/J+/M−/N−) and compounds (KL−/OP+), and later tested how participants would predict the consequences of these cues in novel combinations (e.g., IJ?; MN?) or as singles cues (K? L?; O? P?). The authors observed that a subset of participants showed a generalization pattern consistent with this opposites rule such that they predicted the outcome would occur after MN, K, and L and predicted that it would not occur after IJ, O, and P. This pattern of behavior is hard to reconcile with an associative network which derives its predictions based on feature overlap and thus would predict the exact opposite pattern. Even if knowledge gained about the complete patterning discriminations (A− B− AB+; C+ D+ CD−) is represented within the associative system, it would not be activated in the IJ? or MN? test trials and thus influence the outcome expectation. Maes et al. (2015) have shown that this pattern of abstract rule generalization is absent from the behavior of rats and pigeons, which appear to generalize mainly in ways consistent with associative learning principles. Cobos et al. (2016) showed the same is true for humans when using a cued-response priming task, whereas verbal ratings were consistent with rule-based generalization. Furthermore, the use of rule-based generalization has been shown to be related to working memory, cognitive reflection, and strategic model-based choice in other instrumental learning tasks (Wills et al., 2011a,b; Don et al., 2015, 2016). However, as with the Perruchet effect, researchers have not yet explored whether these competing forms of generalization have an impact on the strength of future learning. Given that several cognitive correlates of rule extraction can be used to predict which individuals are most likely to use a relational rule in this task, predictions can be made about which individuals should find it surprising when a new trial type violates the rule and which should not.

These avenues for future research are among several that might be fruitful for testing how associative predictions and expectations based on non-associative factors might contribute to new learning. Given that most of the current evidence is consistent with multiple theoretical accounts (including those that retain and those that reject classical association formation as a key explanatory construct), devising new experimental designs is essential for the advancement of the field.

CONCLUSION

Having valid and reliable expectations about future events is one of the most essential and necessary conditions for the adaptivity of human behavior. Associative learning theories have offered a very successful account of how humans obtain these expectations and how they update and optimize them whenever these expectations are violated. However, by necessity, formal implementations of these theories in associative networks have a limited scope, which does not capture the influence of a variety of other cognitive factors on our learned judgments and expectations. We have explored three ways how these sources of non-associative knowledge can affect associative learning without changing the fundamental principles of such an associative learning system. We argue that recent theorists have failed to give these possibilities due credence and, even though there is no specific evidence for any of them, they offer plausible ways in which an associative learning and memory system may contribute to judgments and expectations that is consistent with most of the available evidence. Future research is needed to examine whether and how associative predictions and other sources of expectations contribute to future associative learning.

AUTHOR CONTRIBUTIONS

AT and EL were equally responsible for the conception, drafting, and revising of the paper.

FUNDING

The research reported in this article is partly based on work done while AT was a postdoctoral researcher at the University of Sydney, financed by grants from the German Academic Exchange Service (DAAD) and the Australian Research Council (ARC), and while EL was a Mercator Fellow to AT's Project TH 1923/1-1 awarded by the German Research Foundation (DFG). EL's contribution was supported by Australian Research Council grants DP130100864 and DP160102871.

REFERENCES

Annau, Z., and Kamin, L. J. (1961). The conditioned emotional response as a function of intensity of the US. *J. Comp. Physiol. Psychol.* 54, 428–432. doi: 10.1037/h0042199

Barrett, L. C., and Livesey, E. J. (2010). Dissociations between expectancy and performance in simple and two-choice reaction-time tasks: a test of associative and nonassociative explanations. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 864–877. doi: 10.1037/a0019403

Beckers, T., De Houwer, J., Pineno, O., and Miller, R. R. (2005a). Outcome additivity and outcome maximality influence cue competition in human causal learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 238–249.

Beckers, T., Van den Broeck, U. V., Renne, M., Vandorpe, S., Houwer, J. D., and Eelen, P. (2005b). Blocking is sensitive to causal structure in 4-year-old and 8-year-old children. *Exp. Psychol.* 52, 264–271. doi: 10.1027/1618-3169.52.4.264

Blanco, F., Baeyens, F., and Beckers, T. (2014). Blocking in human causal learning is affected by outcome assumptions manipulated through causal structure. *Learn. Behav.* 42, 185–199. doi: 10.3758/s13420-014-0137-y

Blanco, F., Matute, H., and Vadillo, M. A. (2012). Mediating role of activity level in the depressive realism effect. *PLoS ONE* 7:e46203. doi: 10.1371/journal.pone.0046203

Boddez, Y., Haesen, K., Baeyens, F., and Beckers, T. (2014). Selectivity in associative learning: a cognitive stage framework for blocking and cue competition phenomena. *Front. Psychol.* 5:1305. doi: 10.3389/fpsyg.2014.01305

Cobos, P. L., Gutiérrez-Cobo, M. J., Morís, J., and Luque, D. (2016). Dependent measure and time constraints modulate the competition between conflicting feature-based and rule-based generalization processes. *J. Exp. Psychol. Learn. Mem. Cogn.* doi: 10.1037/xlm0000335 [Epub ahead of print].

Colagiuri, B., and Livesey, E. J. (2016). Contextual cuing as a form of nonconscious learning: theoretical and empirical analysis in large and very large samples. *Psychonomic Bull. Rev.* 23, 1996–2009. doi: 10.3758/s13423-016-1063-0

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learn. Behav.* 37, 1–20. doi: 10.3758/LB.37.1.1

De Houwer, J., Beckers, T., and Glautier, S. (2002). Outcome and cue properties modulate blocking. *Q. J. Exp. Psychol.* 55A, 965–985. doi: 10.1080/02724980143000578

Dickinson, A., Shanks, D., and Evenden, J. (1984). Judgement of act-outcome contingency: the role of selective attribution. *Q. J. Exp. Psychol.* 36, 29–50. doi: 10.1080/14640748408401502

Don, H. J., Goldwater, M. B., Otto, R., and Livesey, E. J. (2016). Rule abstraction, model-based choice and cognitive reflection. *Psychonomic Bull. Rev.* 23, 1615–1623.

Don, H. J., Goldwater, M. B., Otto, R. A., and Livesey, E. J. (2015). "Connecting rule-abstraction and model-based choice across disparate learning tasks," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, eds D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, et al. (Pasadena, CA: Cognitive Science Society), 590–595.

Don, H. J., and Livesey, E. J. (2015). Resistance to instructed reversal of the learned predictiveness effect. *Q. J. Exp. Psychol.* 68, 1327–1347. doi: 10.1080/17470218.2014.979212

Fiedler, K., and Juslin, P. (2006). *Information Sampling and Adaptive Cognition*. New York, NY: Cambridge University Press.

Goujon, A., Didierjean, A., and Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends Cogn. Sci.* 19, 524–533. doi: 10.1016/j.tics.2015.07.009

Gredebäck, G., Winman, A., and Juslin, P. (2000). "Rational assessments of covariation and causality," in *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, eds L. R. Gleitman and K. Joshi (Mahwah, NJ: Erlbaum), 190–195.

Harris, J., Livesey, E., Gharaei, S., and Westbrook, F. (2008). Negative patterning is easier than a biconditional discrimination. *J. Exp. Psychol. Anim. Behav. Process.* 34, 494–500. doi: 10.1037/0097-7403.34.4.494

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychol. Rev.* 113, 584–605. doi: 10.1037/0033-295X.113.3.584

Harris, J. A., and Livesey, E. J. (2008). Comparing patterning and biconditional discriminations in humans. *J. Exp. Psychol. Anim. Behav. Process.* 34, 144–154. doi: 10.1037/0097-7403.34.1.144

Harris, J. A., and Livesey, E. J. (2010). An attention-modulated associative network. *Learn. Behav.* 38, 1–26. doi: 10.3758/LB.38.1.1

Haselgrove, M. (2010). Reasoning rats or associative animals? A common-element analysis of the effects of additive and subadditive pretraining on blocking. *J. Exp. Psychol. Anim. Behav. Process.* 36, 296–306.

Kamin, L. J. (1968). "Attention-like processes in classical conditioning," in *Miami Symposium on the Prediction of Behavior: Aversive Stimulation*, ed. M. R. Jones (Miami, FL: University of Miami Press), 9–33.

Lachnit, H., Kinder, A., and Reinhard, G. (2002). Are rules applied in Pavlovian electrodermal conditioning with humans general or outcome specific? *Psychophysiology* 39, 380–387. doi: 10.1017/S0048577201393125

Lachnit, H., Lober, K., Reinhard, G., and Kinder, A. (2001). Evidence for the application of rules in Pavlovian electrodermal conditioning with humans. *Biol. Psychol.* 56, 151–166. doi: 10.1016/S0301-0511(01)00067-9

Lashley, K. S., and Wade, M. (1946). The Pavlovian theory of generalization. *Psychol. Rev.* 53, 72–87. doi: 10.1037/h0059999

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., and Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychol. Bull.* 142, 1111–1140. doi: 10.1037/bul0000064

Lee, J. C., and Livesey, E. J. (2012). Second-order conditioning and conditioned inhibition: influences of speed versus accuracy on human causal learning. *PLoS ONE* 7:e49899. doi: 10.1371/journal.pone.0049899

Lee Cheong Lem, V. A., Harris, J. A., and Livesey, E. J. (2015). Testing the limits of the Perruchet effect in choice response time tasks. *J. Exp. Psychol. Anim. Learn. Cogn.* 41, 385–394. doi: 10.1037/xan0000079

Livesey, E., Lee, J., and Shone, L. (2013). "The relationship between blocking and inference in causal learning," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (COGSCI 2013)*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society).

Livesey, E. J., and Boakes, R. A. (2004). Outcome additivity, elemental processing and blocking in human causality judgements. *Q. J. Exp. Psychol.* 57B, 361–379. doi: 10.1080/02724990444000005

Livesey, E. J., and Costa, D. S. J. (2014). Automaticity and conscious control in single and choice response time versions of the Perruchet effect. *Q. J. Exp. Psychol.* 67, 646–664. doi: 10.1080/17470218.2013.824014

Livesey, E. J., and Harris, J. A. (2008). What are flexible representations? commentary on Melchers, Shanks and Lachnit. *Behav. Process.* 77, 437–439. doi: 10.1016/j.beproc.2007.09.006

Livesey, E. J., and Harris, J. A. (2009). Is there room for simple links in a propositional mind? (Commentary on Mitchell et al.). *Behav. Brain Sci.* 32, 212–213. doi: 10.1017/S0140525X09001010

Livesey, E. J., Thorwart, A., and Harris, J. A. (2011). Comparing positive and negative patterning in human learning. *Q. J. Exp. Psychol.* 64, 2316–2333. doi: 10.1080/17470218.2011.605153

Lovibond, P. F., Been, S. L., Mitchell, C. J., Bouton, M. E., and Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Mem. Cogn.* 31, 133–142. doi: 10.3758/BF03196088

Luque, D., Cobos, P. L., and López, F. J. (2008). Interference between cues requires a causal scenario: favorable evidence for causal reasoning models in learning processes. *Learn. Mot.* 39, 196–208. doi: 10.1016/j.lmot.2007.10.001

Luque, D., Flores, A., and Vadillo, M. A. (2013). Revisiting the role of within-compound associations in cue-interaction phenomena. *Learn. behav.* 41, 61–76. doi: 10.3758/s13420-012-0085-3

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychol. Rev.* 82, 276–298. doi: 10.1037/h0076778

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Anim. Learn. Behav.* 4, 186–192. doi: 10.3758/BF03214033

Maes, E., Filippo, G. D., Inkster, A. B., Lea, S. E. G., Houwer, J. D., D'Hooge, R., et al. (2015). Feature- versus rule-based generalization in rats, pigeons and humans. *Anim. Cogn.* 18, 1267–1284. doi: 10.1007/s10071-015-0895-8

Matute, H. (1996). Illusion of control: detecting response-outcome independence in analytic but not in naturalistic conditions. *Psychol. Sci.* 7, 289–293. doi: 10.1111/j.1467-9280.1996.tb00376.x

Matute, H., Arcediano, F., and Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 182–196. doi: 10.1037/0278-7393.22.1.182

McLaren, I. P. L., and Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Anim. Learn. Behav.* 28, 211–246. doi: 10.3758/BF03200258

McLaren, I. P. L., and Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Anim. Learn. Behav.* 30, 177–200. doi: 10.3758/BF03192828

Melchers, K. G., Shanks, D. R., and Lachnit, H. (2008). Stimulus coding in human associative learning: flexible representations of parts and wholes. *Behav. Process.* 77, 413–427. doi: 10.1016/j.beproc.2007.09.013

Mitchell, C. J., De Houwer, J., and Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behav. Brain Sci.* 32, 183–198. doi: 10.1017/S0140525X09000855

Mitchell, C. J., Griffiths, O., Seeto, J., and Lovibond, P. F. (2012). Attentional mechanisms in learned predictiveness. *J. Exp. Psychol. Anim. Behav. Process.* 38, 191–202. doi: 10.1037/a0027385

Mitchell, C. J., and Lovibond, P. F. (2002). Backward and forward blocking in human electrodermal conditioning: blocking requires an assumption

of outcome additivity. *Q. J. Exp. Psychol.* 55B, 311–329. doi: 10.1080/02724990244000025

Mitchell, C. J., Lovibond, P. F., and Condoleon, M. (2005). Evidence for deductive reasoning in blocking of causal judgments. *Learn. Mot.* 36, 77–87. doi: 10.1016/j.lmot.2004.09.001

Mitchell, C. J., Wardle, S. G., Lovibond, P. F., Weidemann, G., and Chang, B. P. I. (2010). Do reaction times in the Perruchet effect reflect variations in the strength of an associative link? *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 567–572. doi: 10.1037/a0018433

Morís, J., Cobos, P. L., Luque, D., and López, F. J. (2014). Associative repetition priming as a measure of human contingency learning: evidence of forward and backward blocking. *J. Exp. Psychol. Gen.* 143, 77–93. doi: 10.1037/a0030919

Myers, K., Vogel, E., Shin, J., and Wagner, A. (2001). A comparison of the Rescorla-Wagner and Pearce models in a negative patterning and a summation problem. *Anim. Learn. Behav.* 29, 36–45. doi: 10.3758/BF03192814

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford: Oxford University Press.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Anim. Learn. Behav.* 30, 73–95. doi: 10.3758/BF03192911

Pearce, J. M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87, 532–552. doi: 10.1037/0033-295X.87.6.532

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *Pavlov. J. Biol. Sci.* 20, 163–170.

Perruchet, P. (2015). Dissociating conscious expectancies from automatic link formation in associative learning: a review on the so-called Perruchet effect. *J. Exp. Psychol. Anim. Learn. Cogn.* 41, 105–127. doi: 10.1037/xan0000060

Redhead, E. S., and Pearce, J. M. (1995). Stimulus salience and negative patterning. *Q. J. Exp. Psychol.* 48, 67–83.

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory,* eds A. H. Black, and W. F. Prokasy (New York, NY: Appleton Century Crofts), 64–99. doi: 10.1037/a0030892

Shanks, D. R. (2007). Associationism and cognition: human contingency learning at 25. *Q. J. Exp. Psychol.* 60, 291–309. doi: 10.1080/17470210601000581

Shanks, D. R., and Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *J. Exp. Psychol. Anim. Behav. Process.* 24, 405–415.

Shanks, D. R., and Lopez, F. (1996). Causal order does not affect cue selection in human associative learning. *Mem. Cogn.* 24, 511–522. doi: 10.3758/BF03200939

Shone, L. T., Harris, I. M., and Livesey, E. J. (2015). Automaticity and cognitive control in the learned predictiveness effect. *J. Exp. Psychol. Anim. Learn. Cogn.* 41, 18–31. doi: 10.1037/xan0000047

Soto, F. A., Vogel, E. H., Castillo, R. D., and Wagner, A. R. (2009). Generality of the summation effect in human causal learning. *Q. J. Exp. Psychol.* 62, 877–889. doi: 10.1080/17470210802373688

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* 16, 966–973. doi: 10.1038/nn.3413

Sternberg, D. A., and McClelland, J. L. (2012). Two mechanisms of human contingency learning. *Psychol. Sci.* 23, 59–68. doi: 10.1177/0956797611429577

Sutherland, N. S., and Mackintosh, N. J. (1971). *Mechanisms of Animal Discrimination Learning*. New York, NY: Academic Press.

Thorwart, A., and Lachnit, H. (2009). Symmetrical generalization decrements: configural stimulus processing in human contingency learning. *Learn. Behav.* 37, 107–115. doi: 10.3758/LB.37.1.107

Thorwart, A., and Lachnit, H. (2010). Generalization decrements: further support for flexibility in stimulus processing. *Learn. Behav.* 38, 367–373. doi: 10.3758/LB.38.4.367

Thorwart, A., Livesey, E. J., and Harris, J. A. (2012). Normalization between stimulus elements in a model of Pavlovian conditioning: showjumping on an elemental horse. *Learn. Behav.* 40, 334–346. doi: 10.3758/s13420-012-0073-7

Thorwart, A., Uengoer, M., Livesey, E. J., and Harris, J. A. (2016). Summation effects in human learning: evidence from patterning discriminations in goal-tracking. *Q. J. Exp. Psychol.* doi: 10.1080/17470218.2016.1184290 [Epub ahead of print].

Vadillo, M. A., Konstantinidis, E., and Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bull. Rev.* 23, 87–102. doi: 10.3758/s13423-015-0892-6

Vadillo, M. A., and Matute, H. M. (2007). Predictions and causal estimations are not supported by the same associative structure. *Q. J. Exp. Psychol.* 60, 433–447. doi: 10.1080/17470210601002520

Vadillo, M. A., Miller, R. R., and Matute, H. M. (2005). Causal and predictive-value judgments, but not predictions, are based on cue–outcome contingency. *Learn. Behav.* 33, 172–183. doi: 10.3758/BF03196061

Wagner, A. R., and Brandon, S. E. (2001). "A componential theory of Pavlovian conditioning," in *Handbook of Contemporary Learning Theories*, eds R. R. Mowrer and S. B. Klien (Mahwah, NJ: Erlbaum), 23–64.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 53–76. doi: 10.1037/0278-7393.26.1.53

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: evidence from an overshadowing paradigm. *Psychonomic Bull. Rev.* 8, 600–608. doi: 10.3758/BF03196196

Waldmann, M. R., and Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *J. Exp. Psychol. Gen.* 121, 222–236. doi: 10.1037/0096-3445.121.2.222

Weidemann, G., McAndrew, A., Livesey, E. J., and McLaren, I. P. L. (2016). Evidence for multiple processes contributing to the Perruchet effect: Response priming and associative learning. *J. Exp. Psychol. Anim. Learn. Cogn.* 42, 366–379.

Weidemann, G., Tangen, J. M., Lovibond, P. F., and Mitchell, C. J. (2009). Is Perruchet's dissociation between eyeblink conditioned responding and outcome expectancy evidence for two learning systems? *J. Exp. Psychol. Anim. Behav. Process.* 35, 169–176. doi: 10.1037/a0013294

Wills, A. J., Barrasin, T. J., and McLaren, I. P. L. (2011a). "Working memory capacity and generalization in predictive learning," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, eds L. Carlson, C. Hölscher, and T. Shipley (Austin, TX: Cognitive Science Society), 3205–3210.

Wills, A. J., Graham, S., Koh, Z., McLaren, I. P. L., and Rolland, M. D. (2011b). Effects of concurrent load on feature- and rule-based generalization in human contingency learning. *J. Exp. Psychol. Anim. Behav. Process.* 37, 308–316. doi: 10.1037/a0023120

Yarritu, I., Matute, H., and Vadillo, M. A. (2014). Illusion of control: the role of personal involvement. *Exp. Psychol.* 61, 38–47. doi: 10.1027/1618-3169/a000225

Conflict of Interest Statement: 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.

Copyright © 2016 Thorwart and Livesey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

OPINION
published: 24 January 2017
doi: 10.3389/fpsyg.2017.00053

Self-Generated or Cue-Induced—Different Kinds of Expectations to Be Considered

Maike Kemper[1]* *and Robert Gaschler*[2]

[1] *Interdisciplinary Research Cluster on Image, Knowledge, Gestaltung, Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany,* [2] *Interdisciplinary Research Cluster on Image, Knowledge, Gestaltung, Department of Psychology, FernUniversität in Hagen, Hagen, Germany*

Keywords: stimulus expectation, self-generated expectations, predictions, cue-induced expectations, violation of expectation

DIFFERENCES BETWEEN SELF-GENERATED AND CUE-INDUCED EXPECTATIONS

Expectations can help humans to adequately prepare for action. Cognitive psychology has inspired studies on the influence of expectations on the course of scientific discovery (Klein and Roodman, 2005; Rzhetsky et al., 2006; Brewer, 2012). Violations of expectations in research often fail to provoke changes in theorizing and research practices. While expectations have been dissociated from other processes such as automatic response activation (Perruchet et al., 2006), relatively little attention has been devoted to reflecting on the different forms of expectation or different methods used to study expectations (and their violations). In this opinion paper, we highlight some early work (Acosta, 1982) and later contributions that have the potential to violate researchers' expectations on what seems the most suitable methodology for operationalizing expectations in the cognitive psychology lab.

In behavioral research, expectations are most often operationalized by assessing performance differences between trials in which expectations are met and trials in which expectations are violated. Neurophysiological data can assess dynamics before stimulus onset (e.g., Mattler et al., 2006; Kemper et al., 2012), and the mismatch effect shows that response times are faster and error rates lower for expected events (compared to when an expectation is violated). This can be demonstrated for expectations about stimuli (Posner and Snyder, 1975; Mattler, 2004), as well as to-be-performed tasks in task switching studies (Rogers and Monsell, 1995; Meiran, 1996). Furthermore, expectations can modulate the impact of cognitive conflicts (Duthoo et al., 2013, for a study on expectations in the Stroop task). The Gratton effect is a change in the strength of a conflict effect depending on the amount of cognitive conflict in previous trials. It has been described as an expectation effect (Gratton et al., 1992; see also Botvinick et al., 1999; Braver, 2012; but see e.g., Mayr et al., 2003; Schmidt and Weissman, 2016, for alternative interpretations).

Many studies use cues to induce stimulus expectations (Posner and Snyder, 1975; Shulman et al., 1999; Mattler, 2004; Oswal et al., 2007) and task expectations (Rogers and Monsell, 1995; Meiran, 1996). Other methods of inducing expectations include presentation of subliminal stimuli (Kunde, 2004) or irrelevant flankers (Nattkemper et al., 2010; Ziessler and Nattkemper, 2011). At first sight inducing expectations seems to offer a greater degree of experimental control compared to allowing participants to form their own expectations. By inducing expectations, experimenters can determine in advance how often which cue is used and how often the upcoming event violates vs. matches the expectation that the cue should induce. However, studies of stimulus expectations show stronger behavioral (Acosta, 1982) as well as EEG effects (Kemper et al., 2012) for self-generated compared to cue-induced expectations. This suggests that self-generated expectations might nonetheless be preferable, as they induce larger, and therefore more easily detectable, effects.

In a self-generated expectation condition, participants are prompted to verbalize their expectation (e.g., "shape?"). They verbalize which of the stimuli from the set (e.g., "circle") they

OPEN ACCESS

Edited by:
Anna Thorwart,
University of Marburg, Germany

Reviewed by:
Mike Le Pelley,
University of New South Wales, Australia
James R. Schmidt,
Ghent University, Belgium

***Correspondence:**
Maike Kemper
maike.kemper@psychologie.hu-berlin.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 September 2016*
Accepted: *10 January 2017*
Published: *24 January 2017*

Citation:
Kemper M and Gaschler R (2017) Self-Generated or Cue-Induced—Different Kinds of Expectations to Be Considered. Front. Psychol. 8:53. doi: 10.3389/fpsyg.2017.00053

are expecting to appear in the current trial. In a cuing variant, the participants are shown a picture of a circle or the word "circle" (cf. Kemper et al., 2012) and verbalize it. Next, the stimulus is shown and the response is collected. One early example hinting at a qualitative difference between self-generated vs. cue-induced expectations was reported by Acosta (1982). For mismatches, larger stimulus set sizes led to longer response times, for self-generated and cue-induced expectations alike, whereas for match trials, set size effects differed between cue-induced and self-generated expectations. Reaction times for stimuli that matched the cue were longer if more stimuli were used. For self-generated expectations, set size had only a minor influence. Presumably, the expected stimulus was strongly activated regardless of whether there were many or few alternative stimuli. In addition, Acosta (1982) reported evidence that violations of self-generated expectations have a stable effect when prolonging the interval between generation of the expectation and stimulus presentation whereas cue-induced expectations diminish relatively quickly for prolonged intervals. Stronger effects of violations of self-generated compared to cue-induced expectations have not only been obtained for expectations of stimuli. They were also found when expectations concerned a more abstract level of task processing, such as the conflict level of the upcoming trial (e.g., expecting a congruent vs. an incongruent Stroop trial; Kemper et al., 2016). Specifically, expecting the repetition of a congruent trial led to faster processing, while expecting conflict did not enhance performance (e.g., Duthoo et al., 2013). A modulation was found for self-generated expectations only.

Stronger effects of self-generated compared to cue-induced expectations can be attributed to (a) differences in strength and (b) likelihood of engagement. While there is evidence that cues can be ignored (especially in case of low validity; cf. Alpay et al., 2009), even chance-level validity leads to strong effects of self-generated expectations (e.g., Acosta, 1982; Kemper et al., 2012, 2016; Gaschler et al., 2014). This suggests that self-generated expectations cannot be ignored (see Schwager et al., 2016, for a current test of boundary conditions), whereas participants presumably fail to attend to or use cues of low validity in many of the trials. Based on this, and on the lack of a set size effect reported by Acosta (1982), Gaschler et al. (2014) suggested that the object of expectation becomes represented in the focus of attention in working memory (cf. Oberauer et al., 2013) in the case of self-generated expectations (but only occasionally so in the case of cues). This representation is accessible for verbal report, which implies that verbalizations are a rather direct measure of self-generated expectations. The stimulus representation that is activated more strongly than the others (if only by a small margin) can be selected for report.

More specifically, there is evidence for the assumption that this privileged form of representation is a by-product of self-assessing what one is currently expecting. Strong RT benefits for the stimulus that one says one is *not* expecting (Hacker and Hinrichs, 1979) or is expecting second *most* (Hacker and Hinrichs, 1974) suggest that the focus of attention in working memory is filled with the object one considers when self-assessing what one is currently expecting. Thus, researchers should take into account that self-generated verbalized expectations might not only serve as a measure of expectation, but also as a means of boosting expectation effects. Based on the evidence gathered so far, it is difficult to determine whether a stronger effect is in general desirable to increase the internal validity of the experiment or whether this comes at the cost of results that are only representative for the specific situation in which participants are required to form and verbalize an expectation. Studies that directly compare self-generated expectations when participants are either triggered to form them or can form them spontaneously have so far not been conducted. They might become possible using neurophysiological multivariate pattern recognition to trace expectations (cf. Cichy et al., 2014).

RESTRICTED INFLUENCE OF EXPECTATION VIOLATIONS ON SELF-GENERATED EXPECTATIONS

Self-generated expectations allow experimenters to track how violations of expectations influence the formation of future expectations. For example, expectations about the conflict level of an upcoming trial are highly dependent on the conflict level of the previous run of trials. For instance, participants expect a repetition of a conflict trial after one single conflict trial. However, the more conflict trials that have occurred in a row, the stronger participants show a gambler's fallacy and expect a congruent trial next (Jiménez and Méndez, 2013, 2014; Kemper et al., 2016). It is possible that violations of expectations operate like a conflict cue for processing in the upcoming trial. Exploring this possibility might help to further understand the differences between cue-induced and self-generated expectations.

In addition to effects of violation of expectation on future expectations that depend on the last (few) trial(s), stimulus probability influences the overall percentage in which participants predict each stimulus (e.g., Kemper et al., 2012). The phenomenon of probability matching (e.g., Umbach et al., 2012) suggests that self-generated expectations are not strategically chosen to optimize performance. For instance, if Stimulus A is 70% likely and Stimulus B is only 30% likely, probability matching means that participants will anticipate Stimulus A on 70% of the trials, and anticipate Stimulus B on 30% of the trials, even though to optimize performance (by optimizing the number of trials in which expectation and stimulus match) the best solution is to anticipate Stimulus A on 100% of the trials. In principle, participants could exclusively verbalize that they are expecting the frequent stimulus. This would maximize the number of match trials and should improve performance. Such strategic effects would undermine the credibility of verbalizations as a valid measure of expectations. However, participants match their expectations to the probabilities of stimuli instead of minimizing expectation violations (e.g., Kemper et al., 2012; Umbach et al., 2012). Expectations seem to be influenced by and to reflect stimulus frequencies. Future research should explore whether this influence is in part the result of (other) strategic effects. For example, participants might aim to match verbalization frequencies to stimulus frequencies in an attempt to obtain match trials even for the infrequent stimuli. In addition,

probability matching can be an effect of the search for patterns (Gaissmaier and Schooler, 2008).

So far, the evidence for strategic effects is limited. The mismatch effect is stable with practice and of similar strength for frequent and infrequent stimuli, even though violations of expectations are much more likely for infrequent than for frequent stimuli (e.g., Umbach et al., 2012). Frequent violation of an expectation does not influence how strongly that expectation is relied upon in future trials. However, validity might have a different effect on cue induced than on self-generated expectations. Cues show larger mismatch effects when they are relatively valid (i.e., when the expectation is violated less often) and mismatch effects become very small for cues with low validity (e.g., Vossel et al., 2006).

CONCLUSIONS FOR RESEARCH ON (THE VIOLATION OF) EXPECTATIONS

Self-generated expectations show stronger effects than cue-induced expectations in a number of experimental setups, and measure expectations more effectively relative to cues. People still rely on self-generated expectations even if they are violated often (e.g., are of low validity in a long experiment). We suggest that researchers should take into account that the choice between self-generated and cue-induced expectations entails a tradeoff between the strength of the expectation effect and the degree of experimental control over expectations in individual trials. In addition, since internally-generated expectations may differ qualitatively from those induced by cues, it cannot be taken for granted that results obtained with one method can be generalized to situations involving the other.

AUTHOR CONTRIBUTIONS

MK and RG have written this article together. They have both contributed substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the Interdisciplinary Research Cluster Image, Knowledge, Gestaltung (EXC1027/1), at Humboldt-Universität zu Berlin, Berlin.

REFERENCES

Acosta, E. (1982). Subjective and objective determinants of expectancy: similarities and differences. *Am. J. Psychol.* 95, 139–160. doi: 10.2307/1422664

Alpay, G., Goerke, M., and Stürmer, B. (2009). Precueing imminent conflict does not override sequence-dependent interference adaptation. *Psychol. Res.* 73, 803–816. doi: 10.1007/s00426-008-0196-9

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., and Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* 402, 179–181. doi: 10.1038/46035

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends Cogn. Sci. (Regul. Ed).* 16, 106–113. doi: 10.1016/j.tics.2011.12.010

Brewer, F. W. (2012). "The theory ladenness of the mental processes used in the scientific enterprise," in *Psychology of Science: Implicit and Explicit Processes,* eds R. W. Proctor and J. E. Capaldi (New York, NY: Oxford University Press), 289–334. doi: 10.1093/acprof:oso/9780199753628.003.0013

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Duthoo, W., Wühr, P., and Notebaert, W. (2013). The hot-hand fallacy in cognitive control: repetition expectancy modulates the congruency sequence effect. *Psychon. Bull. Rev.* 20, 798–805. doi: 10.3758/s13423-013-0390-7

Gaissmaier, W., and Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition* 109, 416–422. doi: 10.1016/j.cognition.2008.09.007

Gaschler, R., Schwager, S., Umbach, V. J., Frensch, P. A., and Schubert, T. (2014). Expectation mismatch: differences between self-generated and cue-induced expectations. *Neurosci. Biobehav. Rev.* 46, 139–157. doi: 10.1016/j.neubiorev.2014.06.009

Gratton, G., Coles, M. G., and Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *J. Exp. Psychol. Gen.* 121:480. doi: 10.1037/0096-3445.121.4.480

Hacker, M. J., and Hinrichs, J. V. (1974). Multiple predictions in choice reaction time: a serial memory scanning interpretation. *J. Exp. Psychol.* 103, 999–1005. doi: 10.1037/h0037396

Hacker, M. J., and Hinrichs, J. V. (1979). Verbal predictions of unexpected stimuli and choice reaction time. *Mem. Cogn.* 7, 505–510. doi: 10.3758/BF03198267

Jiménez, L., and Méndez, A. (2013). It is not what you expect: dissociating conflict adaptation from expectancies in a Stroop task. *J. Exp. Psychol. Hum. Percept. Perform.* 39:271. doi: 10.1037/a0027734

Jiménez, L., and Méndez, A. (2014). Even with time, conflict adaptation is not made of expectancies. *Front. Psychol.* 5:1042. doi: 10.3389/fpsyg.2014.01042

Kemper, M., Gaschler, R., Schwager, S., and Schubert, T. (2016). The benefit of expecting no conflict—stronger influence of self-generated than cue-induced conflict expectations on Stroop performance. *Acta Psychol. (Amst.).* 163, 135–141. doi: 10.1016/j.actpsy.2015.11.008

Kemper, M., Umbach, V. J., Schwager, S., Gaschler, R., Frensch, P. A., and Stürmer, B. (2012). What I say is what I get: stronger effects of self-generated vs. cue-induced expectations in event-related potentials. *Front. Psychol.* 3:562. doi: 10.3389/fpsyg.2012.00562

Klein, J. R., and Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annu. Rev. Nucl. Particle Sci.* 55, 141–163. doi: 10.1146/annurev.nucl.55.090704.151521

Kunde, W. (2004). Response priming by supraliminal and subliminal action effects. *Psychol. Res.* 68, 91–96. doi: 10.1007/s00426-003-0147-4

Mattler, U. (2004). Combined expectancy effects are modulated by the relation between expectancy cues. *Q. J. Exp. Psychol. A* 57, 193–221. doi: 10.1080/02724980343000161

Mattler, U., van der Lugt, A., and Münte, T. F. (2006). Combined expectancies: electrophysiological evidence for adjusted expectancy effects. *BMC Neurosci.* 7:37. doi: 10.1186/1471-2202-7-37

Mayr, U., Awh, E., and Laurey, P. (2003). Conflict adaptation effects in the absence of executive control. *Nat. Neurosci.* 6, 450–452. doi: 10.1038/nn1051

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *J. Exp. Psychol. Learn. Mem. Cogn.* 22:1423. doi: 10.1037/0278-7393.22.6.1423

Nattkemper, D., Ziessler, M., and Frensch, P. A. (2010). Binding in voluntary action control. *Neurosci. Biobehav. Rev.* 34, 1092–1101. doi: 10.1016/j.neubiorev.2009.12.013

Oberauer, K., Souza, A. S., Druey, M. D., and Gade, M. (2013). Analogous mechanisms of selection and updating in declarative and procedural working memory: experiments and a computational model. *Cogn. Psychol.* 66, 157–211. doi: 10.1016/j.cogpsych.2012.11.001

Oswal, A., Ogden, M., and Carpenter, R. H. (2007). The time course of stimulus expectation in a saccadic decision task. *J. Neurophysiol.* 97, 2722–2730. doi: 10.1152/jn.01238.2006

Perruchet, P., Cleeremans, A., and Destrebecqz, A. (2006). Dissociating the effects of automatic activation and explicit expectancy on reaction times in a simple associative learning task. *J. Exp. Psychol. Learn. Mem. Cogn.* 32:955. doi: 10.1037/0278-7393.32.5.955

Posner, M. I., and Snyder, C. R. R. (1975). "Facilitation and inhibition in the processing of signals," in *Attention and Performance V*, ed P. M. A. Rabbit (London: Academic Press), 669–682.

Rogers, R. D., and Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* 124:207. doi: 10.1037/0096-3445.124.2.207

Rzhetsky, A., Iossifov, I., Loh, J. M., and White, K. P. (2006). Microparadigms: chains of collective reasoning in publications about molecular interactions. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4940–4945. doi: 10.1073/pnas.0600591103

Schmidt, J. R., and Weissman, D. H. (2016). Congruency sequence effects and previous response times: conflict adaptation or temporal learning? *Psychol. Res.* 80, 590–607. doi: 10.1007/s00426-015-0681-x

Schwager, S., Gaschler, R., Rünger, D., and Frensch, P. A. (2016). Tied to expectations: predicting features speeds processing even under adverse circumstances. *Mem. Cogn.* doi: 10.3758/s13421-016-0683-y. [Epub ahead of print].

Shulman, G. L., Ollinger, J. M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Petersen, S. E., et al. (1999). Areas involved in encoding and applying directional expectations to moving objects. *J. Neurosci.* 19, 9480–9496.

Umbach, V. J., Schwager, S., Frensch, P. A., and Gaschler, R. (2012). Does explicit expectation really affect preparation? *Front. Psychol.* 3:378. doi: 10.3389/fpsyg.2012.00378

Vossel, S., Thiel, C. M., and Fink, G. R. (2006). Cue validity modulates the neural correlates of covert endogenous orienting of attention in parietal and frontal cortex. *Neuroimage* 32, 1257–1264. doi: 10.1016/j.neuroimage.2006.05.019

Ziessler, M., and Nattkemper, D. (2011). The temporal dynamics of effect anticipation incourse of action planning. *Q. J. Exp. Psychol.* 64, 1305–1326. doi: 10.1080/17470218.2011.553067

Conflict of Interest Statement: 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.

Copyright © 2017 Kemper and Gaschler. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 December 2016
doi: 10.3389/fpsyg.2016.01968

Reminder Cues Modulate the Renewal Effect in Human Predictive Learning

Javier Bustamante[1], Metin Uengoer[2]* and Harald Lachnit[2]*

[1] Department of Psychology, University of Chile, Santiago, Chile, [2] Faculty of Psychology, Philipps-Universität Marburg, Marburg, Germany

Associative learning refers to our ability to learn about regularities in our environment. When a stimulus is repeatedly followed by a specific outcome, we learn to expect the outcome in the presence of the stimulus. We are also able to modify established expectations in the face of disconfirming information (the stimulus is no longer followed by the outcome). Both the change of environmental regularities and the related processes of adaptation are referred to as extinction. However, extinction does not erase the initially acquired expectations. For instance, following successful extinction, the initially learned expectations can recover when there is a context change – a phenomenon called the renewal effect, which is considered as a model for relapse after exposure therapy. Renewal was found to be modulated by reminder cues of acquisition and extinction. However, the mechanisms underlying the effectiveness of reminder cues are not well understood. The aim of the present study was to investigate the impact of reminder cues on renewal in the field of human predictive learning. Experiment I demonstrated that renewal in human predictive learning is modulated by cues related to acquisition or extinction. Initially, participants received pairings of a stimulus and an outcome in one context. These stimulus-outcome pairings were preceded by presentations of a reminder cue (acquisition cue). Then, participants received extinction in a different context in which presentations of the stimulus were no longer followed by the outcome. These extinction trials were preceded by a second reminder cue (extinction cue). During a final phase conducted in a third context, participants showed stronger expectations of the outcome in the presence of the stimulus when testing was accompanied by the acquisition cue compared to the extinction cue. Experiment II tested an explanation of the reminder cue effect in terms of simple cue-outcome associations. Therefore, acquisition and extinction cues were equated for their associative histories in Experiment II, which should abolish their impact on renewal if based on simple cue-outcome associations. In contrast to this prediction, Experiment II replicated the findings from Experiment I indicating that the effectiveness of reminder cues did not require direct reminder cue-outcome associations.

Keywords: human learning, extinction, renewal, context, retrieval cue

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Gonzalo P. Urcelay, University of Leicester, UK
Shira Meir Drexler, Ruhr University Bochum, Germany

****Correspondence:***
Javier Bustamante
je.bustamante@gmail.com
Metin Uengoer
uengoer@staff.uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *23 May 2016*
Accepted: *02 December 2016*
Published: *20 December 2016*

Citation:
Bustamante J, Uengoer M and Lachnit H (2016) Reminder Cues Modulate the Renewal Effect in Human Predictive Learning. Front. Psychol. 7:1968. doi: 10.3389/fpsyg.2016.01968

INTRODUCTION

Background stimuli play a relevant role in the behavioral expression of learning. Extinction performance, for instance, seems to be particularly vulnerable to context changes (Bouton, 2004; Urcelay and Miller, 2014), as shown by the renewal effect. In a typical renewal procedure, a conditioned stimulus (CS; e.g., tone) is repeatedly paired with an unconditioned stimulus (US; e.g., shock) in Context A establishing conditioned responding (CR; e.g., fear) toward the CS. Then during extinction, the CS is presented repeatedly alone in Context B, which causes a gradual reduction in the response level elicited by the CS. Finally, when the participants are tested again in the acquisition Context A, the originally learned behavior reappears. This recovery effect is referred to as ABA renewal, with the letters denoting the contexts of acquisition, extinction, and test. Renewal has also been reported when acquisition, extinction, and testing take place in three different contexts (ABC renewal; Bouton and Bolles, 1979), and when acquisition and extinction are conducted in the same context and testing in a different one (AAB renewal; Bouton and Ricker, 1994). The renewal effect is a cardinal example for the persistence of expectations in the face of disconfirming information. The initially acquired expectations are not erased but suppressed instead by extinction. But this suppression is highly context-specific (Bouton, 1993, 2004).

The renewal effect is also considered as a model for relapse after exposure-based treatments (Bouton, 2000; Bouton et al., 2006). In exposure therapy, a patient is confronted with a problematic stimulus in order to decrease responding to it, for example, by exposing a phobic patient to the fear-eliciting event or stimulus. The renewal effect indicates that the therapeutic success in overcoming unwanted responses will be linked to a certain degree to the therapeutic environment. When a patient leaves the treatment context, relapse is facilitated.

Different strategies to influence the strength of renewal have been examined in the conditioning literature (for a review, see Laborda et al., 2011; Craske et al., 2014). One of these treatments is the use of reminder cues. For example, using a human fear conditioning task, Vansteenwegen et al. (2006) demonstrated that ABA renewal was affected by a reminder cue (a black cross) correlated with either acquisition or extinction. In one group, the reminder cue preceded the trials during the acquisition phase conducted in Context A, while in a second group the cue preceded the trials during extinction in Context B. Finally, all participants received presentations of the cue during a test of response recovery in Context A. Vansteenwegen et al. (2006) observed stronger renewal in those participants for which the cue was previously trained during initial acquisition than in those for which the cue previously accompanied extinction. Furthermore, the ability of reminder cues to modulate response recovery has been documented in a variety of preparations, including appetitive conditioning (Brooks and Bouton, 1994; Brooks, 2000; Brooks and Bowker, 2001) and ethanol tolerance (Brooks et al., 1999) in rats as well as fear conditioning (Dibbets et al., 2008; Dibbets and Maes, 2011), fear of spiders (Dibbets et al., 2013), and reactivity to alcohol-signaling cues (e.g., Collins and Brandon, 2002) in humans.

The aim of the present study was to extend the results of Vansteenwegen et al. (2006) to human predictive learning (Experiment I), and to examine a potential mechanism that may underlie the modulatory impact of reminder cues on response recovery (Experiment II). According to Brooks and Bouton (1993, 1994), there is the possibility that reminder cues might act through direct cue-outcome associations (e.g., Rescorla and Wagner, 1972). This view assumes that a cue presented in close temporal proximity to reinforcement of a CS acquires excitatory associative strength, while a reminder cue presented during extinction develops an inhibitory cue-outcome association. This view received support from a human fear conditioning experiment by Dibbets and Maes (2011) who observed that a cue presented during extinction of one CS attenuated conditioned responding to a second CS (summation test; Rescorla, 1969) indicating that the extinction reminder cue directly inhibited the US-representation.

Other studies, however, have shown that the effectiveness of reminder cues can be independent of any direct associations with the outcome. For example, it has been reported that an extinction reminder cue reduced response recovery even though it did not pass a summation test for conditioned inhibition (Brooks and Bouton, 1993; Dibbets et al., 2008). Furthermore, Brooks and Bowker (2001) showed that an extinction reminder cue still decreased response recovery after being paired with the US.

Experiment I was aimed at replicating the modulatory impact of acquisition and extinction reminder cues on response recovery reported by Vansteenwegen et al. (2006) for fear conditioning to human predictive learning, using a task with an ABC renewal procedure. Experiment II examined the importance of direct cue-outcome associations for the effectiveness of reminder cues. Therefore, we used an experimental design in which the acquisition and extinction reminder cues were equated for their associative histories. Each reminder cue was followed by the outcome on half of the trials, and was presented without the outcome on the other half. If the effectiveness of reminder cues relies on direct associations with the outcome, this treatment should abolish the impact of the cues on renewal. Both experiments were implemented in a predictive learning task that asked participants to imagine being a medical doctor whose patient often suffers from stomach trouble after the consumption of different meals in different restaurants (e.g., Üngör and Lachnit, 2006). The task was to predict the occurrence (+) or non-occurrence (−) of this stomach trouble. On successive trials, different stimuli (food types) were presented in one of several contexts (restaurants), and participants were asked to predict the patient's reaction. On trials with a reminder cue, each food/restaurant presentation was preceded by a brief presentation of a picture showing either a cup of coffee or a glass of wine. During the learning phases of each experiment, each trial ended with information about whether stomach trouble had occurred or not.

EXPERIMENT I

Table 1 illustrates the design for the two groups of Experiment I. During Phase 1, all participants received Z+ trials in Context A (acquisition), with 80% of the trials preceded by a reminder cue (Y). During Phase 2, participants received training with Z− in Context B (extinction), with 80% of the trials preceded by a second reminder cue (X). Finally, during Phase 3 (Test) participants received trials with Z in Contexts B and C. For half of the participants (Group AC – acquisition cue) each of the test trials in Context C was preceded by the reminder Cue Y, the one presented during the acquisition phase, while for the other half of participants (Group EC – extinction cue) the trials in Context C were preceded by the reminder Cue X from the extinction phase. Thus, the Test consisted of an ABC renewal procedure, and each group was tested with a reminder cue correlated with either acquisition or extinction. If the reminder cues exert influence on responding during Test, we should find a lower level of renewal in Group EC than in Group AC.

Method

Participants

The participants were 46 students from the Philipps-Universität Marburg, Germany (33 women and 13 men). Their age varied between 17 and 29 years, with a median of 22. They either were paid (€1.50), rewarded with chocolate or received course credits for participation. Participants were equally allocated to the different experimental groups as they arrived in the experimental room. They were tested individually and required between 10 and 15 min to complete the experiment. The data of 19 additional participants were excluded from the analyses because their predictions were incorrect on more than 30% of the trials with Stimulus Z during the last two blocks in Phase 1 and/or during the last two blocks in Phase 2. All participants gave their written consent to participate in the experiment.

Apparatus and Procedure

Instructions and all necessary information were presented on a computer screen. Participants interacted with the computer using the mouse. The following food types were used as stimuli: apples, avocados, bananas, broccoli, eggs, strawberries, carrots, corn, tomatoes, grapes, and lemons. The pictures of a glass of red wine and a cup of coffee were used as reminder cues. The names of three fictitious restaurants were used as contexts, labeled (translated from German) "To The Mug," "By The Innkeeper," and "In The Kettle," written in red, blue, and green font, respectively. The assignment of the different food types to Stimulus Z and Filler Cues F1–F10 as well as the assignment of the restaurant names to the contexts were randomized for each participant. The pictures of the glass of wine and the cup of coffee were also randomly assigned to the acquisition and extinction cues. During the learning phases, each trial ended with either the presentation of the outcome (+; occurrence of stomach troubles) or with its absence (−; non-occurrence of stomach troubles).

Initially, each participant was asked to read the instructions (complete instructions attached as "Supplementary Material"). They were instructed to imagine being a medical doctor, and that one of their patients suffers frequently of stomach troubles after meals. Participants were told that their patient goes out often for meals to some restaurants. After each visit to a restaurant the participant would have to predict whether the patient suffers of stomach troubles or not.

Each trial started with a blank screen with a gray background presented for 500 ms followed by the name of one of the restaurants surrounded by a rectangular frame of the color associated with the restaurant. On trials with a reminder cue, in addition the picture of either a glass of wine or a cup of coffee was presented on the center of the screen. After 1000 ms, a picture of one food type replaced the reminder cue if it was present. The name of the food was written below the picture. Participants were told that their patient had eaten the food at the restaurant. They were instructed to make a prediction of whether they expect that their patient suffers from stomach troubles. Participants made their predictions by clicking on one of two answer buttons labeled "Yes, I expect stomach trouble," and "No, I do not expect stomach trouble," which were located below the food picture. Immediately after participants responded, another window appeared, telling the participants whether their patient suffered of stomach troubles or not. Participants had to confirm that they had read the feedback by clicking on an "OK" button. Then the next trial started.

During Phase 1 (see **Table 1**), all participants were given 10 trials of Z+ and F1− each in Context A, 10 trials of F2+, and F3− each in Context B, and 10 trials of F4+ and F5− each in Context C. The acquisition reminder Cue Y was present in 8 of the Z+ trials; the trials in which the reminder cue was shown were determined randomly. In Phase 2, all participants received 10 trials of F6+ and F7− each in Context A, 10 trials of Z− and F8− each in Context B, and 10 trials of F9+ and F10− each in Context C. The second reminder Cue X preceded Z− in 8 of the trials, assigned randomly. Trials with Stimulus Z in Phase 1 and Phase 2 that were not preceded by a reminder cue ensured that participants already experienced this stimulus in the absence of reminder cues prior to the Test (see below; see also, Brooks and Bouton, 1993, 1994; Vansteenwegen et al., 2006). Phase 2 followed Phase 1 without a break (the transition was not signaled to the participants).

Phase 1 and Phase 2 each were divided into five blocks, with each block consisting of two presentations of each

TABLE 1 | A summary of the experimental design of Experiment I (A, B, and C represent different restaurant names; Stimulus Z refers to the picture of a food item; Cues Y and X are pictures of two different drinks; + and − are occurrence and non-occurrence of stomach troubles, respectively; ?, participants received no feedback; the experimental design comprised additional filler cues that are not depicted in the table – see "Method" Section for details).

Group	Phase 1	Phase 2	Test
AC	A: Y/Z+	B: X/Z−	B: Z? C: Y/Z?
EC	A:Y/Z+	B: X/Z−	B: Z? C: X/Z?

food stimulus. The order of presentation of the trials within each block was determined randomly for each block and participant.

Phase 3 (Test) was introduced by instructions to the participants informing that the feedback would be omitted, but that they should try to predict the occurrence or non-occurrence of the outcome (complete instructions as "Supplementary Material"). Test trials were identical to learning trials, with the exception that the feedback window was omitted. All participants were presented with four Z trials in Context B and four trials with Z in Context C. For half of the participants (Group AC) each trial with Z in Context C was preceded by the acquisition Cue Y, whereas for the other half (Group EC) these trials were preceded by the extinction Cue X. The Test was divided into two blocks, and within each block each trial type was presented two times. The order of presentation of the trials within each block was determined randomly.

Results

For this and the subsequent experiment, the 0.05 level of significance was employed for all statistical tests, and stated probability levels were based on the Greenhouse and Geisser (1959) adjustment of degrees of freedom where appropriate (for the sake of readability, we report uncorrected degrees of freedom). We report partial eta squared (η_p^2) as the measure of effect size.

Acquisition (Phase 1)

The left-hand panel of **Figure 1** presents for each group the mean percentages of stomach trouble predictions for Z+ in Context A across the five blocks of Phase 1. Black squares represent the data of Group AC, and white squares the data of Group EC. As can be seen, the mean prediction to Z+ increased across blocks, and there were no differences in responding to Z+ between groups. This was confirmed by a 5 × 2 (Block [1, 2, 3, 4, 5] × Group [AC, EC]) ANOVA. A significant main effect of Block was found, $F(4,176) = 23.11$, $p < 0.001$, $\eta_p^2 = 0.344$, indicating an increase of stomach trouble predictions to Z+ over the course of acquisition training, but neither a significant main effect of Group nor a significant Block × Group interaction was detected, all $Fs < 1$, showing that there was no difference in the predictions between groups.

Extinction (Phase 2)

The right-hand panel of **Figure 1** presents for each group the mean percentages of stomach trouble predictions for Z− in Context B across the five blocks of Phase 2. As depicted in **Figure 1**, the mean of stomach trouble predictions decreased across blocks, showing that the response to Z was successfully extinguished. This was confirmed by a 5 × 2 (Block [1, 2, 3, 4, 5] × Group [AC, EC]) ANOVA. There was a significant main effect of Block, $F(4,176) = 54.40$, $p < 0.001$, $\eta_p^2 = 0.553$, but no significant main effect of Group, $F(1,44) = 1.78$, $p = 0.188$, $\eta_p^2 = 0.039$, and no significant Block × Group interaction, $F < 1$, were detected, confirming that there were no differences between groups.

Test

Figure 2 depicts responding to Z in Contexts B and C during the Test in terms of the mean percentages of stomach trouble predictions, collapsed across the four test trials presented in each context. The left-hand bars present the predictions for Group AC and the right-hand bars show the predictions for Group EC.

As **Figure 2** demonstrates, participants in Group AC showed a higher level of responding to Z in Context C than in Context B (ABC renewal), while participants in Group EC showed similar levels of responding across the two contexts, indicating an absence of response recovery due to context changes. A 2 × 2 (Context [B, C] × Group [AC, EC]) ANOVA revealed a significant main effect of Context, $F(1,44) = 12.38$, $p < 0.002$, $\eta_p^2 = 0.22$, a significant main effect of Group, $F(1,44) = 7.57$, $p < 0.009$, $\eta_p^2 = 0.147$, and most importantly, a significant Context × Group interaction, $F(1,44) = 22.24$, $p < 0.001$, $\eta_p^2 = 0.336$, indicating that context-dependency of responding was stronger in Group AC than in Group EC. Further analyses were conducted on each group to explore the Context × Group interaction. A paired-samples t-test in Group AC yielded significantly stronger responding to Z in Context C than in Context B, $t(22) = 6.45$, $p < 0.001$, while there was no such a difference in Group EC, $t < 1$. These comparisons confirmed the presence and absence of renewal in Group AC and Group EC, respectively.

Discussion

Taken together, after acquisition and extinction were conducted in two different contexts, testing the target stimulus in a third context disrupted extinction performance (ABC renewal) only if the test trials were preceded by a reminder cue related to initial acquisition training. When the test trials were preceded by a reminder cue related to extinction learning, however, extinction performance generalized perfectly to the third context.

The present results replicate the findings reported in human fear conditioning by Vansteenwegen et al. (2006) using an ABA procedure. The present results extend their findings to a human predictive learning procedure without biologically significant stimuli as well as to an ABC renewal design, both demonstrating the generality of the previous work.

In the learning phases of the present experiment, presentations of the acquisition reminder Cue Y were always followed by the outcome (occurrence of stomach trouble), while trials with the extinction reminder Cue X were consistently followed by its absence (non-occurrence of stomach trouble). When presented during Test, Y and X might have retrieved memories of their related outcomes which encouraged the participants to predict stomach trouble when the target stimulus was preceded by Y, and to predict its absence when the target was preceded by X. The purpose of the following experiment was to test this explanation in terms of direct reminder cue-outcome associations.

EXPERIMENT II

Table 2 depicts the design for the two groups of Experiment II. The learning and test phases were identical to those of

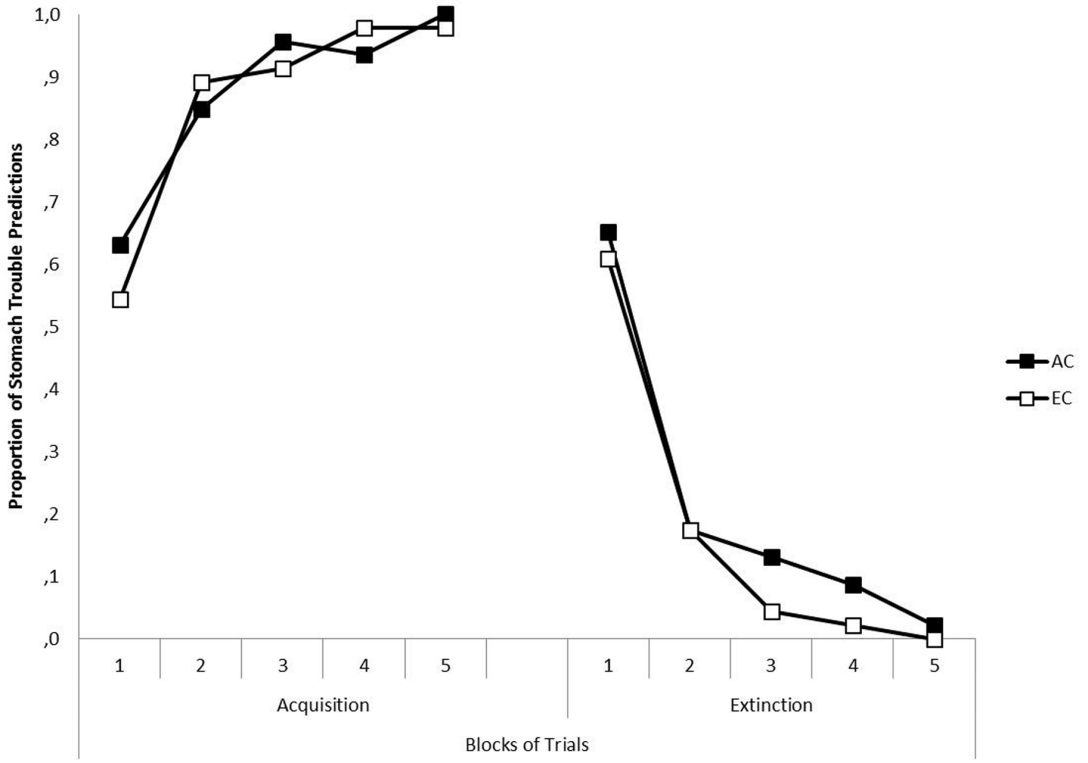


FIGURE 1 | The left-hand panel shows the mean proportion of predictions of stomach trouble in response to Z in Context A across five blocks in the acquisition phase of Experiment I, separately for Groups AC (black squares) and EC (white squares). The right-hand panel shows the mean proportion of predictions of stomach trouble in response to Z in Context B across five blocks in the extinction phase of Experiment I for Groups AC and EC.

Experiment I, with the exceptions that the acquisition reminder Cue Y additionally preceded 80% of the trials with F3− in Context B during Phase 1, and that the extinction reminder Cue X also preceded 80% of the trials with F6+ in Context A during Phase 2. Thus, in Experiment 2, acquisition and extinction reminder cues were equated for their learning histories in the way that each reminder cue was associated with the outcome on half of its presentations, while on the other half it was followed by the absence of the outcome. If reminder cues influence performance during the Test by retrieving memories related to their associated outcomes, then we should observe no difference in response recovery across the two groups in the present experiment.

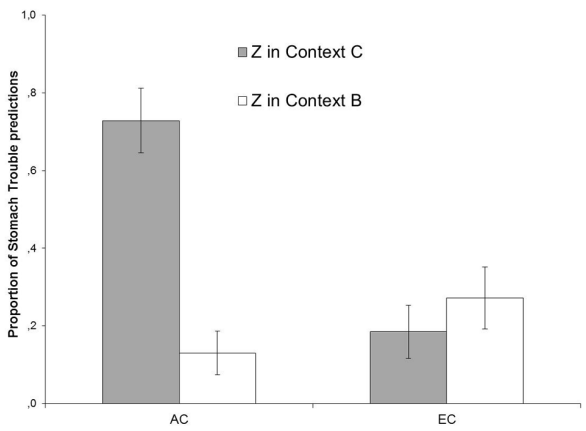


FIGURE 2 | Mean proportions of predictions of stomach trouble in response to Z during the test phase of Experiment I, collapsed across four presentations of each trial type separately for Groups AC and EC in Contexts B and C. Error bars denote standard error of the means.

Method

Participants, Apparatus, and Procedure

The participants were 58 students from the Philipps-Universität Marburg, Germany (29 women and 29 men). Their age varied between 19 and 49 years, with a median of 22. The data of 21 additional participants were excluded from the analyses because their predictions were incorrect on more than 30% of the trials with Stimulus Z during the last two blocks in Phase 1 and/or during the last two blocks in Phase 2. All participants gave their written consent to participate in the experiment. The stimuli, instructions and procedure of Experiment II were the same as those of Experiment I, with the exceptions that the acquisition reminder Cue Y also preceded 8 of the 10 trials with F3− in Context B during Phase 1, and that the extinction reminder Cue X also preceded 8 of the 10 trials with F6+ in Context A during Phase 2. For each of the Stimuli F3 and F6, the trials in which the reminder cue was shown were determined randomly.

TABLE 2 | A summary of the experimental design of Experiment II (A, B, and C represent different restaurant names; Stimuli Z, F3, and F6 refer to pictures of different food items; Cues Y and X are pictures of two different drinks; + and − are occurrence and non-occurrence of stomach troubles, respectively; ?, participants received no feedback; the experimental design comprised additional filler cues that are not depicted in the table – see "Method" Section for details).

Group	Phase 1	Phase 2	Test
AC	A: Y/Z+	B: X/Z−	B: Z?
	B: Y/F3−	A: X/F6+	C: Y/Z?
EC	A:Y/Z+	B: X/Z−	B: Z?
	B: Y/F3−	A: X/F6+	C: X/Z?

Results

Acquisition (Phase 1)

The left-hand panel of **Figure 3** presents for each group the mean percentages of stomach trouble predictions for Z+ in Context A across the five blocks of Phase 1. Black squares represent the data of Group AC, and white squares the data of Group EC. As can be seen, the mean prediction to Z+ increased across blocks, and there were no differences in responding to Z+ between groups. This was confirmed by a 5 × 2 (Block [1, 2, 3, 4, 5] × Group [AC, EC]) ANOVA. A significant main effect of Block was found, $F(4,224) = 33.68$, $p < 0.001$, $\eta_P^2 = 0.376$, indicating an increase of stomach trouble predictions to Z+ over the course of acquisition training, but neither a significant main effect of Group nor a significant Block × Group interaction was detected, both $Fs < 1$, showing that there was no difference in the prediction levels between groups.

Extinction (Phase 2)

The right-hand panel of **Figure 3** presents for each group the mean percentages of stomach trouble predictions for Z− in Context B across the five blocks of Phase 2. As depicted, the means of stomach trouble predictions decreased across blocks, showing that the response to Z was successfully extinguished. This was confirmed by a 5 × 2 (Block [1, 2, 3, 4, 5] × Group [AC, EC]) ANOVA. There was a significant main effect of Block, $F(4,224) = 77.57$, $p < 0.001$, $\eta_P^2 = 0.581$, but neither a significant main effect of Group nor a significant Block × Group interaction was detected, both $Fs < 1$, confirming that there were no differences between groups.

Test

Figure 4 depicts responding to Z in Contexts B and C during the Test in terms of the mean percentages of stomach trouble predictions, collapsed across the four test trials presented in each context. The left-hand bars present the predictions for Group AC, and the right-hand bars show the predictions for Group EC.

As **Figure 4** demonstrates, Group AC showed a higher level of responding to Z in Context C than in Context B, while Group EC showed similar levels of responding across the two contexts. A 2 × 2 (Context [B, C] × Group [AC, EC]) ANOVA revealed no significant main effect of Context, $F(1,56) = 1.55$, $p = 0.218$, $\eta_P^2 = 0.027$, no significant main effect of Group, $F(1,56) = 2.11$, $p = 0.15$, $\eta_P^2 = 0.036$, but there was a significant Context × Group interaction, $F(1,56) = 12.09$, $p < 0.001$, $\eta_P^2 = 0.178$, indicating that context-dependency of responding was stronger in Group AC than in Group EC. Paired-samples t-tests showed that participants in Group AC responded significantly stronger to Z in Context C than in Context B, $t(28) = 3.35$, $p < 0.002$, whereas there was no such difference in Group EC, $t(28) = 1.57$, $p = 0.127$.

Discussion

The results from the Test of Experiment II were the same as those from Experiment I. Participants showed ABC renewal when testing occurred in the presence of a cue that had been experienced during initial acquisition learning. However, extinction performance was not disrupted by contextual changes when testing took place in the presence of a cue that had been administered during extinction treatment. In Experiment II, the two reminder cues did not differ with respect to their association with the outcome. Each reminder cue was paired with the outcome on half of its presentations. Thus, the modulation does not require direct reminder cue-outcome associations.

GENERAL DISCUSSION

In two human predictive learning experiments, we observed stronger response recovery following extinction when test trials were preceded by a reminder cue of initial acquisition compared to testing in the presence of an extinction reminder cue. Additionally, in Experiment II the acquisition and extinction cues were equated for their associative histories. Each reminder cue was followed by the outcome on half of the trials, indicating that the effect of the reminder cues does not require direct reminder cue-outcome associations.

Our study extends the generality of the conclusion drawn from previous experiments that the effect of a reminder cue can be independent of a direct association between the reminder cue and the outcome. Brooks and Bouton (1994) and Dibbets et al. (2008) found no evidence that an extinction reminder cue acquired inhibitory associative strength. Brooks and Bowker (2001) reported that an extinction cue did not lose its modulatory impact after being paired with the US. Our study is the first to provide evidence for this conclusion in a human predictive learning paradigm using an ABC renewal protocol. By equating the associative histories of the reminder cues, we extend the scope of methods demonstrating that the effectiveness of reminder cues is not necessarily a function of their own schedule of reinforcement.

Our results are rather consistent with the view that reminder cues modulate retrieval of entire CS–US associations akin to occasion setters (Holland, 1983, 1989; Rescorla, 1986; Schmajuk and Holland, 1998). An alternative explanation for the present results is provided by configural learning theories (Pearce, 1987, 1994). According to this view, the specific reminder cue-CS pattern might be encoded as a unique representation which would develop a direct connection to the US-representation. Future research might aim to differentiate

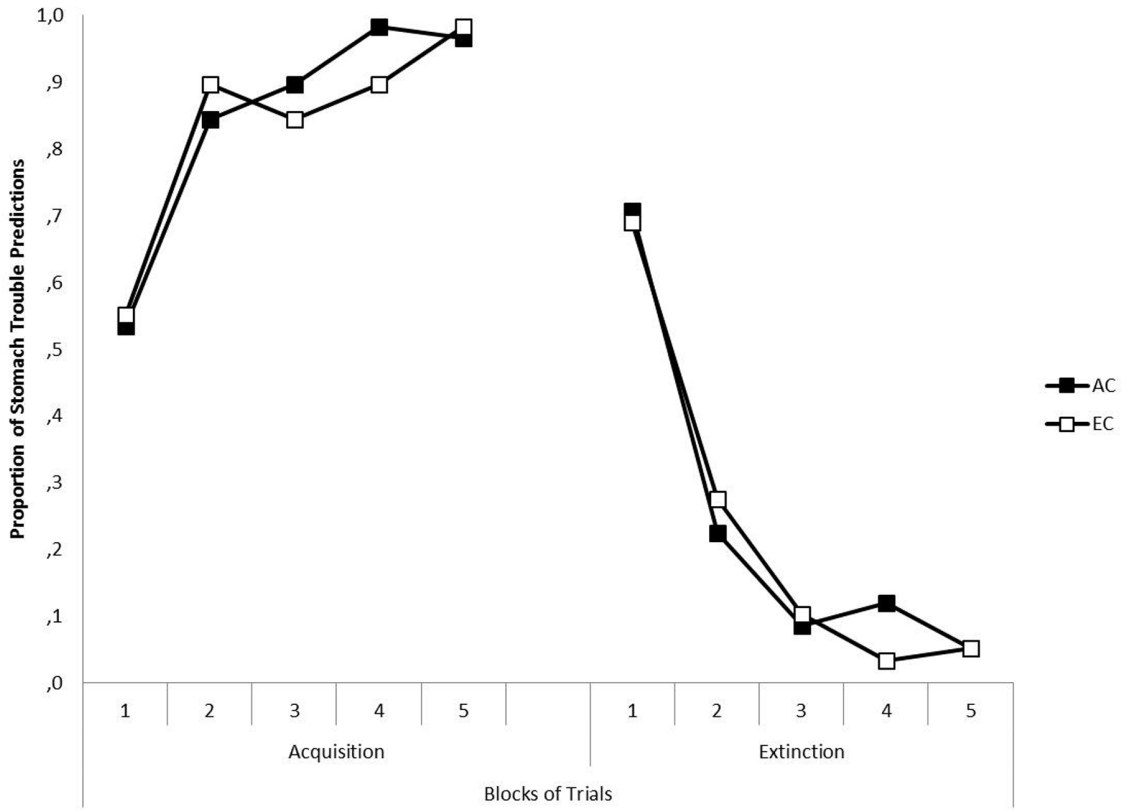


FIGURE 3 | The left-hand panel shows the mean proportion of predictions of stomach trouble in response to Z in Context A across five blocks in the acquisition phase of Experiment II, separately for Groups AC (black squares) and EC (white squares). The right-hand panel shows the mean proportion of predictions of stomach trouble in response to Z in Context B across five blocks in the extinction phase of Experiment II for Groups AC and EC.

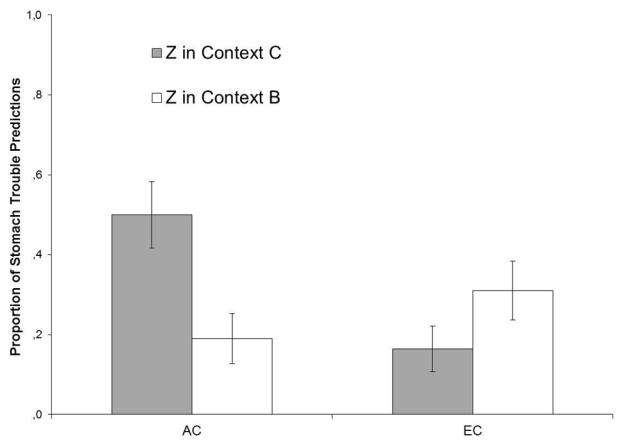


FIGURE 4 | Mean proportions of predictions of stomach trouble in response to Z during the test phase of Experiment II, collapsed across four presentations of each trial type separately for Groups AC and EC in Contexts B and C. Error bars denote standard error of the means.

between the configural and the occasion setting hypotheses, for example, by examining whether a reminder cue shows transfer of its modulatory properties to a second CS with an inconsistent reinforcement history, but not to other stimuli that were consistently paired with an outcome. This selective transfer is a hallmark of occasion setting (Holland, 1989) which cannot be explained by standard configural theories (Pearce, 1987, 1994).

The idea that reminder cues influence performance through their direct connections to the outcome cannot explain the results from our second experiment. However, this account provides a straightforward explanation of Experiment I. Therefore, we cannot exclude the possibility that reminder cue-outcome associations at least contributed to the recovery effects in the present study. In fact, there is some evidence for such a contribution when cross-experimental comparisons are taken into account. We observed stronger ABC renewal in Group AC from Experiment I than in Group AC from Experiment II. This was confirmed by a 2 × 2 (Context [B, C] × Group [AC/Experiment I, AC/Experiment II]) ANOVA revealing a Context × Group interaction, $F(1,50) = 4.69$, $p = 0.035$, $\eta_P^2 = 0.086$. This finding could be explained by assuming that the acquisition reminder cue in Experiment I acquired stronger excitatory strength than the one in Experiment II. However, we found no evidence for a contribution of direct cue-outcome associations in case of the extinction reminder cue. A 2 × 2 (Context [B, C] × Group [EC/Experiment I, EC/Experiment II])

ANOVA revealed no Context × Group interaction, $F < 1$. This latter finding is inconsistent with our analysis, but might also be considered to reflect a floor effect. Thus, the direct associations account could at least explain aspects of our data. However, conclusions from cross-experimental comparisons should be treated with caution, and future research will be required to investigate possible contributions of reminder cue-outcome associations to the strength of response recovery.

Our understanding of the mechanisms underlying the effectiveness of reminder cues has important implications for a clinical application (Craske et al., 2014). For instance, if an extinction reminder cue supports retrieval of the inhibitory CS–US association, this cue can be used as a powerful tool to enhance the long-term success of exposure-based treatments. However, if an extinction cue acts through a direct inhibitory connection to the US, then the cue should be removed from the clinical setting as it would be detrimental to the therapeutic goals. In this case, the cue would be a "safety signal," for instance, signaling the absence of fear which would protect the fear-eliciting target stimulus from extinction.

In two experiments, we show that reminder cues exerted influence on the strength of response recovery following extinction in a predictive learning task. However, our experiments were not designed to assess the individual contributions of acquisition and extinction reminder cues to this behavioral modulation. The difference in response recovery during the test phase of each experiment might have been caused by (a) an increase of renewal due to the presentation of the acquisition cue, (b) a decrease in renewal by the extinction cue, or (c) both (see also Vansteenwegen et al., 2006). However, in each of our experiments, response recovery was completely abolished when testing was conducted in the presence of the extinction cue. Taken into account studies using a similar procedure demonstrating robust ABC renewal in the absence of reminder cues (e.g., Üngör and Lachnit, 2008), this diminution can be considered as indirect evidence that the extinction cue contributed to performance by reducing response recovery. However, future research is required to test this directly and to disentangle the individual and relative contributions of acquisition and extinction reminder cues on response recovery.

ETHICS STATEMENT

The experimental procedure was approved by the ethics committee of the Psychology Department of the Philipps-Universitaet Marburg. Before the experiment, participants were informed about the general purpose and the general procedure of the experiment. And they were briefed that they can cancel their participation at any point during the experiment. Participants gave informed written consent to participate in the experiment. They confirmed their participation as voluntary and agreed to the use of their data in anonymous form for scientific purposes.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

JB was supported by a doctoral scholarship from the German Academic Exchange Service (Deutscher Akademischer Austauschdienst-DAAD) and by a postdoctoral grant from the Fondo Nacional de Desarrollo Científico y Tecnológico (Postdoctoral Fondecyt #3160591).

ACKNOWLEDGMENTS

We thank Kathrin Bahlinger, Joanna Buryn-Weizel, Dominik Deffner, Barnd Hengstebeck, Lukas Herbst, Jascha Kristek, Anne-Marie Leonhardt, Simon Samstag, and Francisco Wilhelm for their help with data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01968/full#supplementary-material

REFERENCES

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigm of Pavlovian learning. *Psychol. Bull.* 114, 80–99. doi: 10.1037/0033-2909.114.1.80

Bouton, M. E. (2000). A learning theory perspective on lapse, relapse, and the maintenance of behavior change. *Health Psychol.* 19, 57–63. doi: 10.1037/0278-6133.19.Suppl1.57

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learn. Mem.* 11, 485–494. doi: 10.1101/lm.78804

Bouton, M. E., and Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learn. Motiv.* 10, 445–466. doi: 10.1016/0023-9690(79)90057-2

Bouton, M. E., and Ricker, S. T. (1994). Renewal of extinguished responding in a second context. *Anim. Learn. Behav.* 22, 317–324. doi: 10.3758/BF03209840

Bouton, M. E., Woods, A. M., Moody, E. W., Sunsay, C., and García-Gutiérrez, A. (2006). "Counteracting the context-dependence of extinction: relapse and tests of some relapse prevention methods," in *Fear and Learning: From Basic Processes to Clinical Implications*, eds M. G. Craske, D. Hermans, and D. Vansteenwegen (Washington, DC: American Psychological Association), 175–196.

Brooks, D. C. (2000). Recent and remote extinction cues reduce spontaneous recovery. *Q. J. Exp. Psychol. B* 53, 25–58. doi: 10.1080/027249900392986

Brooks, D. C., and Bouton, M. E. (1993). A retrieval cue for extinction attenuates spontaneous recovery. *J. Exp. Psychol. Anim. Behav. Process.* 19, 77–89.

Brooks, D. C., and Bouton, M. E. (1994). A retrieval cue for extinction attenuates response recovery (renewal) caused by a return to the conditioning context. *J. Exp. Psychol. Anim. Behav. Process.* 20, 366–379.

Brooks, D. C., and Bowker, J. L. (2001). Further evidence that conditioned inhibition is not the mechanism of an extinction cue's effect: a reinforced cue

prevents spontaneous recovery. *Anim. Learn. Behav.* 29, 381–388. doi: 10.3758/BF03192903

Brooks, D. C., Palmatier, M., Garcia, E. O., and Johnson, J. L. (1999). An extinction cue reduces spontaneous recovery of a conditioned taste aversion. *Anim. Learn. Behav.* 27, 77–88. doi: 10.3758/BF03199433

Collins, B. N., and Brandon, T. H. (2002). Effects of extinction context and retrieval cues on alcohol cue reactivity among nonalcoholic drinkers. *J. Consult. Clin. Psychol.* 70, 390–397. doi: 10.1037/0022-006X.70.2.390

Craske, M., Treanor, M., Conway, C., Zbozinek, T., and Vervliet, B. (2014). Maximizing exposure therapy: an inhibitory learning approach. *Behav. Res. Ther.* 58, 10–23. doi: 10.1016/j.brat.2014.04.006

Dibbets, P., Havermans, R., and Arntz, A. (2008). All we need is a cue to remember: the effect of an extinction cue on renewal. *Behav. Res. Ther.* 46, 1070–1077. doi: 10.1016/j.brat.2008.05.007

Dibbets, P., and Maes, J. H. R. (2011). The effect of an extinction cue on ABA-renewal: Does valence matter? *Learn. Motiv.* 42, 133–144. doi: 10.1016/j.lmot.2010.12.003

Dibbets, P., Moor, C., and Voncken, M. J. (2013). The effect of a retrieval cue on the return of spider fear. *J. Behav. Ther. Exp. Psychiatry* 44, 361–367. doi: 10.1016/j.jbtep.2013.03.005

Greenhouse, S. W., and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* 24, 95–112. doi: 10.1007/BF02289823

Holland, P. C. (1983). "Occasion-setting in Pavlovian feature positive discriminations," in *Quantitative Analyses of Behavior: Discrimination Processes*, eds M. L. Commons, R. J. Herrnstein, and A. R. Wagner (New York, NY: Ballinger), 183–206.

Holland, P. C. (1989). Transfer of negative occasion setting and conditioned inhibition across conditioned and unconditioned stimuli. *J. Exp. Psychol. Anim. Behav. Process.* 15, 311–328.

Laborda, M. A., McConnell, B. L., and Miller, R. R. (2011). "Behavioral techniques to reduce relapse after exposure therapy: applications of studies of experimental extinction," in *Associative Learning and Conditioning Theory: Human and Non-Human Applications*, eds T. R. Schachtman and S. Reilly (New York, NY: Oxford University Press), 79–103.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychol. Rev.* 94, 61–73. doi: 10.1037/0033-295X.94.1.61

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychol. Rev.* 101, 587–607. doi: 10.1037/0033-295X.101.4.587

Rescorla, R. A. (1969). Pavlovian conditioned inhibition. *Psychol. Bull.* 72, 77–94. doi: 10.1037/h0027760

Rescorla, R. A. (1986). Extinction of facilitation. *J. Exp. Psychol. Anim. Behav. Process.* 12, 16–24.

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement," in *Classical Conditioning II: Current Theory and Research*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century Crofts), 64–99.

Schmajuk, N. A., and Holland, P. C. (1998). *Occasion Setting, Associative Learning and Cognition in Animals*. Washington, DC: American Psychological Association.

Üngör, M., and Lachnit, H. (2006). Contextual control in discrimination reversal learning. *J. Exp. Psychol. Anim. Behav. Process.* 32, 441–453.

Üngör, M., and Lachnit, H. (2008). Dissociations among ABA, ABC, and AAB Recovery Effects. *Learn. Motiv.* 39, 181–195. doi: 10.1016/j.lmot.2007.08.001

Urcelay, G. P., and Miller, R. R. (2014). The functions of contexts in associative learning. *Behav. Process.* 104, 2–12. doi: 10.1016/j.beproc.2014.02.008

Vansteenwegen, D., Vervliet, B., Hermans, D., Beckers, T., Baeyens, F., and Eelen, P. (2006). Stronger renewal in human fear conditioning when tested with an acquisition retrieval cue than with an extinction retrieval cue. *Behav. Res. Ther.* 44, 1717–1725. doi: 10.1016/j.brat.2005.10.014

Conflict of Interest Statement: 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.

Copyright © 2016 Bustamante, Uengoer and Lachnit. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 February 2017
doi: 10.3389/fpsyg.2017.00120

Compound Stimulus Presentation Does Not Deepen Extinction in Human Causal Learning

Oren Griffiths, Nathan Holmes and R. Fred Westbrook*

School of Psychology, University of New South Wales, Sydney, NSW, Australia

Models of associative learning have proposed that cue-outcome learning critically depends on the degree of prediction error encountered during training. Two experiments examined the role of error-driven extinction learning in a human causal learning task. Target cues underwent extinction in the presence of additional cues, which differed in the degree to which they predicted the outcome, thereby manipulating outcome expectancy and, in the absence of any change in reinforcement, prediction error. These prediction error manipulations have each been shown to modulate extinction learning in aversive conditioning studies. While both manipulations resulted in increased prediction error during training, neither enhanced extinction in the present human learning task (one manipulation resulted in less extinction at test). The results are discussed with reference to the types of associations that are regulated by prediction error, the types of error terms involved in their regulation, and how these interact with parameters involved in training.

Keywords: extinction, Pavlovian conditioning, animal conditioning, human learning, prediction error

OPEN ACCESS

Edited by:
Mario Gollwitzer,
University of Marburg, Germany

Reviewed by:
Tom Joseph Barry,
King's College London, UK
Bridget L. McConnell,
James Cook University Singapore, Singapore

****Correspondence:***
Oren Griffiths
oren.griffiths@unsw.edu.au

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 September 2016*
Accepted: *17 January 2017*
Published: *09 February 2017*

Citation:
Griffiths O, Holmes N and Westbrook RF (2017) Compound Stimulus Presentation Does Not Deepen Extinction in Human Causal Learning. Front. Psychol. 8:120. doi: 10.3389/fpsyg.2017.00120

INTRODUCTION

Prediction error refers to the degree of mismatch between what is expected to occur, and what actually occurs. One way to elicit a prediction error is an extinction procedure. In this procedure subjects (animals or people) are first exposed to pairings of a cue (labeled A) and an outcome (denoted +). As a consequence of having experienced several of these A+ pairings, subjects begin to respond to the cue in anticipation of the outcome. This is referred to as the acquisition phase. It is after this acquisition phase that the crucial extinction phase takes place. In the extinction phase, the cue is repeatedly presented in the absence of the outcome, referred to as A− trials. On each of these trials, the expectation of the outcome (+) elicited by the presence of the cue (A) is violated by the experimenter withholding that outcome After several errors in prediction (or prediction errors) whereby the outcome is anticipated but fails to occur, subjects learn that the cue no longer signals the outcome, and responses to the cue cease. At this point the cue-outcome association is said to be extinguished. Thus extinction learning is said to be error-driven, as it is the experience of this prediction error that drives the changes in expectation of the outcome following the cue, and thus elicits learning about that cue. The processes underlying the extinction of such responses are of theoretical and clinical significance, as extinction learning is the basis of exposure therapy, the most effective treatment for many anxiety disorders (Feske and Chambless, 1995; Taylor, 1996; Eddy et al., 2004). Therefore any procedure that purports to enhance extinction learning offers the prospect of enhancing the efficacy of its real-world applications, such as exposure therapy.

The present experiments investigated a method recently reported to enhance extinction learning (and thus exposure therapy) in adults via increasing the prediction error term on each extinction

trial (Culver et al., 2015). Specifically, we tested whether two such manipulations influenced explicit extinction learning in an adult population using affectively neutral stimuli. Neutral stimuli were used in order to focus on the basic cue-outcome learning processes involved in the manipulation, and to correspondingly minimize any differential directs effects that, say, electric shock or its omission might have on learning. In order to understand these error-enhancing manipulations, it is important to first consider *how* prediction error is thought to drive learning more generally.

A range of experiments show that the formation of cue-outcome associations is regulated by prediction error. Specifically, these experiments show that the amount learned about a cue depends not only on its relation to the outcome stimulus, but also on the relation between other concomitantly present cues and that outcome. For example, the "blocking" effect demonstrated that pairings of a target cue (A) with the outcome (+), which would otherwise lead to strong learning about the relationship between the cue and the outcome, could be rendered ineffective by changing which other cues were present on that same trial. For example, if cue A was also accompanied by a second cue (B) that had been previously been trained to predict the outcome, thus rendering cue A causally redundant, then very little is learned about cue A's relationship with the outcome; this is termed the "blocking" effect (Kamin, 1969). In prediction error terms, on the crucial compound trials (AB+ trials), the outcome (+) was already predicted by the second cue (B), and thus there was no prediction error present to drive learning about the target cue (A). Several related empirical phenomena support the role of error-correction mechanisms in acquisition learning in both animals (conditioned inhibition, Rescorla, 1969; overshadowing, Rescorla, 1970; signal validity effects, Wagner, 1969) and people (conditioned inhibition, Chapman and Robbins, 1990; blocking, Dickinson et al., 1984; super-conditioning, Aitken et al., 2000).

There is evidence from animal conditioning studies that extinction learning is also regulated by prediction error. For example, in both between- and within-subject designs, Leung et al. (2012) extinguished one target cue (A) in compound with a partner (X) that was strongly associated with the outcome, and a second target cue (B) in compound with a partner (Y) that was only weakly associated with the outcome. Thus, there was greater prediction error on AX− than on BY− trials, but the treatment of the target cues (A and B) was otherwise identical. The subsequent test of A and B revealed less conditioned responses to A, extinguished in compound with the strong associate of the outcome, X, than to B, extinguished in compound with the weak associate of the outcome, Y. The larger error across the AX− than the BY− trials increased the amount of extinction learning to A than B (see also Leung and Westbrook, 2008; Holmes and Westbrook, 2013). However, there is also evidence from animal conditioning studies that does not suggest that extinction learning depends on the size of the prediction error term. McConnell et al. (2013) used a between-group design to compare the amount of extinction learning to a target conditioned stimulus non-reinforced in compound with either two neutral cues, one neutral cue and one conditioned cue, or two conditioned cues. They found mixed evidence regarding whether extinction learning is driven by the size of the prediction error term. Consistent with the view the extinction learning is driven by prediction error magnitude, they reported that a target conditioned stimulus elicited less responding at test (more extinction) if it had been non-reinforced in compound with one neutral and one conditioned cue than in compound with two neutral cues. Yet they also reported that a target conditioned stimulus elicited less responding at test if it had been non-reinforced in compound with one neutral and one conditioned cue than in compound with two conditioned cues, suggesting that extinction learning is not just controlled by the size of the error term (see also Pearce and Wilson, 1991; Thomas and Ayres, 2004; Witnauer and Miller, 2012).

Recent studies have examined whether evidence for deepened extinction observed by Leung et al. (2012) and others (Leung and Westbrook, 2008, 2010; Holmes and Westbrook, 2013) can also be found in people. Three of these studies used an aversive conditioning procedure in which the experimenters measured both skin conductance levels and the degree to which participants expected an aversive outcome following presentation of the cue. One reason for using both measures is that skin conductance, but not expectancy, is thought to reveal implicit "non-conscious" learning (McAndrew et al., 2012) (but see Mitchell et al., 2009). The first study (Lovibond et al., 2000) examined whether extinction was greater to an excitor (a cue paired with shock) extinguished in conjunction with another excitor (prediction error was large) than to an excitor extinguished in compound with a learned safety signal (prediction error was small). However, there was no such difference on test: each of the target excitors elicited similar levels of test responding (on both measures), suggesting that the cues had failed to interact in the manner expected based on results from animal conditioning studies. In the second study, Vervliet et al. (2007) reported that extinguishing an excitor in compound with a second excitor resulted in performance at test (on both measures) comparable with pre-extinction levels of fear, suggesting that the second excitor had not only failed to enhance extinction learning about the first but had even protected the first from extinction. Again, this result suggests that the cues had failed to interact in the expected manner when presented in compound.

The third study, Culver et al. (2015), offers the most direct test of the proposal that error-correction mechanisms regulate extinction. In this between-groups study, an excitor was subjected to an initial phase of extinction, and then additional extinction either on its own or in compound with a current excitor. This was the method used by Leung et al. (2012), as it is under these conditions that many error-correction theories unambiguously predict a deepening of extinction in the compound group. Consistent with the findings reported by Leung et al. (2012), Culver et al. (2015) found that extinction in the presence of the current excitor deepened extinction of the skin conductance response: this was evidenced by greater resistance to reinstatement of such responses following exposure to the aversive event alone in the group submitted to compound extinction than in the group submitted to further extinction of the target alone. However, as in the two other aversive conditioning studies, Culver et al. (2015) failed to detect any effect of compound extinction on expectancy ratings.

Cumulatively, the literature shows that, at least under some conditions, error correction mechanisms regulate extinction of affective reactions to cues predictive of aversive events in both animals (Leung et al., 2012) and people (Culver et al., 2015). However, at present, there is no evidence that these same mechanisms regulate extinction of the explicit cue-outcome contingency in people. Whether contingency knowledge is regulated by an error-correction process remains an important question to address as cognitive factors have been shown to play a critical role in human extinction learning (for a review, see Lovibond, 2004). For example, Zeng et al. (2015; see also Raes et al., 2011) recently demonstrated that, once a cue-outcome relationship is successfully extinguished, fear of that cue can be immediately restored by providing an alternative explanation for the absence of the aversive outcome during the extinction training. That is, if people reappraise the extinction experiments as providing no evidence about the status of the underlying cue-outcome relationship (akin to using "safety behaviors" in a clinical setting; Salkovskis, 1991), then their fear of the extinguished cue is restored. This observation is consistent with the common sense notion that understanding the cause of aversive events critically influences subsequent behavioral and emotional responses (e.g., Clark, 1986). Similarly, extinction can also be rendered less effective if people aggregate across their whole experience with a cue (when it signals an aversive event in acquisition, and when it signals no such event in extinction), rather than prioritize their most recent experiences with that cue (i.e., during extinction; Collins and Shanks, 2002). Both of these phenomena, reappraisal and aggregation, indicate that effective learning will depend on how people formulate the change in the relation between cues and outcomes across extinction training.

Accordingly, the present study examined the effect of extinction on people's knowledge of the relations between affectively neutral cues and outcomes. It specifically examined whether extinction of a target cue-outcome relationship is regulated by prediction error, which was manipulated through the associative status of cues that accompanied the target during extinction. Across both experiments, steps were undertaken to investigate the role (if any) of aggregation. Specifically, additional "filler" cues were included to assess whether people were aggregating their experiences with cues across phases when asked to assess those cues at test. Moreover, the wording of each test question was adjusted from prior experiments (e.g., Griffiths and Westbrook, 2012; Holmes et al., 2014) to indicate that people should rely on their recent experience with a cue, rather than their remote experience. However, the primary aim of Experiment 1 was to address whether extinction was directly regulated by a prediction error term, using a design analogous to those used by Lovibond et al. (2000) and Vervliet et al. (2007; see also Reberg, 1972). The target cue was extinguished in compound with a good predictor of the outcome (thus eliciting a large prediction error during extinction) while a second cue was extinguished in compound with an already-extinguished cue (thus eliciting a smaller prediction error during extinction). Experiment 2 addressed the same question with a design analogous to that used by Culver et al. (2015). The already-extinguished target cue was given further extinction in compound with another already-extinguished cue – a manipulation that has been shown to restore responding and deepen extinction learning in animal conditioning studies (Hendry, 1982; Rescorla, 2006; Leung et al., 2012). The effects of this compound extinction were assessed relative to a second cue given further extinction in isolation. If extinction learning is regulated by prediction error, the target cue in each experiment should undergo more learning than the control cue, evoking a weaker expectancy of the outcome than the control cue at test.

EXPERIMENT 1

Both experiments used an allergist task, which is a common method for studying associative learning in people (Aitken et al., 2000; Griffiths and Mitchell, 2009). In this task, participants are asked to monitor the intake and symptoms of a fictional patient (in this case, Mrs. X) who suffers from food allergies. The foods the patient consumes are the cues, and any allergic reactions she has are the outcomes. Learning about Mrs. X's food allergies essentially constitutes learning about cue-outcome associations in a trial-by-trial manner. Participants were additionally told that Mrs. X was undergoing chemotherapy, and that her food allergies may consequently vary across time. The design of Experiment 1 is shown in the upper row of **Table 1**. Four foods, e.g., carrots, beef, apples, pasta, labeled as cues A, B, C, and D, are of major interest. Other foods are also presented in each training phase as

TABLE 1 | Experimental design of Experiments 1 and 2 (in the top and middle row, respectively).

	Phase 1 (8)	Phase 2 (8)	Phase 3 (8)	Forced choice
Experiment 1	A++	A–	AB–	B vs. D
	B++			A vs. C
	C++		CD–	
	D++			
	E–	E++		
	F–	F–		
Experiment 2	A++	A–	AB–	B vs. D
	B++	B–		A vs. C
	C++	C–	C–	
	D++	D–	D–	
	E–	E++		
	F–	F–		
Distracters (Both Experiment 1 and Experiment 2)	G–	G–	GH++	
	H–	H–		
	I–	I+		
	J+	J+	J+	
		K+		

Each letter (A–K) refers to an individual food cue (e.g., carrot). The symbols (–,+,++) refer to the severity of the allergic reaction experienced by Mrs. X on each trial: "–" refers to no reaction, "+" refers to minor reaction, "++" refers to a serious reaction. The distracter items listed in the lower row were common to both experiments. The numbers in brackets in the header row refer to the number of trials per trial-type in each Phase.

so-called filler cues. The manipulation of interest occurs in Phase 3 when one allergenic food (B) is extinguished in compound with a food (A) already known to be safe (AB−), whereas a second allergenic food (C) is extinguished in compound with D another allergenic food (CD−). The meals with two allergenic foods present (CD− trials) should elicit more prediction error, and drive more extinction learning for those foods (C and D), than should the meals which contain only one allergenic food (AB− trials). More precisely, the shift from A− to AB− should deepen extinction of A and protect B from extinction (Rescorla, 2006; Leung et al., 2012), while extinction of the compound containing the two allergenic foods (CD−) should be rapid and substantial. Error correction theories thus predict that these manipulations will have contrasting effects on extinction: A will protect B from extinction whereas C will facilitate extinction to D (as will D facilitate extinction of C). According to such theories, therefore, participants will judge B as less safe (or more allergenic) than C (and D) at test. We tested their knowledge of the cue-outcome associations (food-allergy associations) with forced choice items and confidence ratings for each cue.

Materials and Methods

This study was approved by the UNSW Human Research Ethics Advisory Panel, and was carried out in accordance with the recommendations of the National Health and Medical Research Committee's National Statement on Ethical Conduct in Human Research.

Participants

Sixty eight second-year psychology students participated in partial fulfillment of course requirements. The mean age was 19.41 years ($SD = 4.50$), and 45 were female.

Design

The experiment involved three training phases followed by test. In both this and the subsequent experiment, the critical cues are labeled A–D. The remaining cues (E–K) were included to control for any relatively simple, incidental rule learning that might occur (e.g., no meal of two foods produces an allergic reaction). We did not attempt to control for more complex rules (e.g., negative patterning) for the simple reason that people view such complex cue interactions as inherently implausible (Griffiths et al., 2009).

Accordingly, our description of the training contingencies focuses on cues A–D. In Phase 1, each of these four cues (A, B, C, D) was paired with a serious allergic reaction (labeled ++). Other cues were paired with either a mild reaction (labeled +) or no allergic reaction (labeled −).

In Phase 2, one of the previously allergenic cues was extinguished (A−). In Phase 3, two compounds (AB− and CD−) were extinguished. Each of these compounds contained a cue that still predicted an allergic reaction, B and D. However, the status of its partner cue within that compound differed. B was paired with the already extinguished A, whereas D was paired with another allergenic cue, C. Therefore, the prediction error elicited by compounds AB and CD will differ, such that more prediction error will be evoked on CD− than AB− trials. Correspondingly, there will be more extinction learning on CD− than on AB− trials. The filler cues, E–K, were selected so as to balance the number of compounds that did or did not cause allergic reactions and that were followed by allergic reactions in each phase. These cues also balanced the number of cues that changed their relation to the allergenic reaction between phases, and the number of cues presented in isolation or in compound.

Measures

There were four dependent variables: outcome predictions, confidence ratings, test ratings and forced choice responses. The first two occurred during the training phases, and the latter two occurred during the test phase. An outcome prediction was made on every training trial, following the presentation of the cue stimuli. These predictions were made using an onscreen "antibody scale" that varied between 0.0 and 6.0 in 0.1 increments (see **Figure 1**). The scale was visually divided into three categories: no reaction (0–2), minor reaction (2.1–4), and serious reaction (4.1–6). This scale was present on every trial. Participants were told that this scale indicated Mrs. X's "antibody levels (a measure of allergic reaction severity)" after eating each food (see Griffiths et al., 2011). Each time they moved the scrollbar to make a prediction, the numeric value of the scrollbar (e.g., 1.6) was shown on-screen, as was the category of reaction (none, minor, serious) that corresponded to that prediction.

The confidence scale consisted of a five-point scale (where 1 was "not at all confident" and 5 was "very confident"), whereby people rated their confidence in each outcome prediction. The scale was shown in the lower portion of the screen following each outcome prediction rating (see **Figure 1**).

The test ratings were made individually for each food cue on separate screens. At the initiation of this test phase, participants were told that Mrs. X was undergoing medical treatment and therefore had to keep her antibody levels within the normal range. Consequently, the participant had to identify which foods were or were not safe for Mrs. X to eat right now by rating each food on a 0.0–6.0 scale (where 0 was "very unlikely to produce a reaction next time she eats it" and 6 was "very likely to produce a reaction next time she eats it"). Like the scale anchors, the wording of the test question on the screen ("how likely is this food to produce an allergic reaction in Mrs. X right now?") and also the wording of the instructions for this test phase both emphasized the importance of the participants rating the *current* status of the food, rather than providing a rating based on averaging over the history of their experience with that food (see Collins and Shanks, 2002).

On each forced choice test item, participants were shown two meals (which each consisted of a single food), and they were asked to click on the meal that would be safer for Mrs. X to eat at that moment. The left/right positions of each food cue was randomly determined for each participant.

Procedure

The experiment was conducted in classes of approximately 20 students per class. The task was computer-based. Participants were first instructed to assume the role of an allergist who had to learn which foods made a new patient (Mrs. X) feel ill and those which were safe for her to consume.

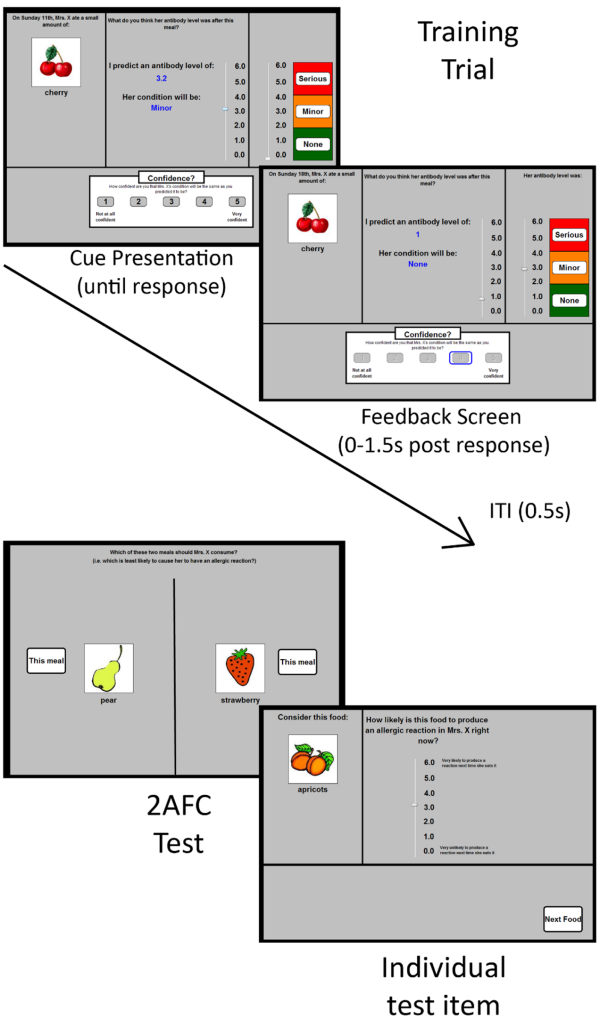


FIGURE 1 | Schematic depiction of the materials used in Experiments 1 and 2. The **(top left)** panel shows a typical training trial in which participants were shown a food (or two foods, not shown) and were first asked to predict Mrs. X's resultant antibody levels. They did this by manipulating the scrollbar in the panel. When that prediction was made, the participant was asked to rate their confidence using the 5-option response scale at the bottom of the screen. Feedback was then provided immediately for 1.5 s. A typical feedback screen is shown in the **(top right)** panel. It differs from the cue presentation screen in that it shows the correct value (on the right-hand scrollbar), alongside the prediction and confidence values chosen by the participant (left hand scrollbar and blue rectangle on the lower screen, respectively). Mrs. X's antibody levels (and thus her allergic reaction response) was indicated by the right-hand response scale. The participant's chosen confidence response remained onscreen. An intertrial interval (ITI) of 0.5 s occurred between trials, during which the preceding trial's cue, response information, and feedback was removed from the screen. The lower two panels depict the two types of test-items. The **(lower left)** panel shows a typical forced choice test-item. The **(lower right)** item shows a typical test item in which people were asked to rate the allergenic properties of each food item individually.

On each trial, participants were shown a meal containing either a single food (e.g., the word "carrots" and a color line drawing of carrots) or two foods (e.g., "beans and broccoli" and a line drawing of each). They were asked to predict whether consumption of the meal would cause an allergenic reaction (see **Figure 1**) using the outcome prediction scale (see *Measures*). Foods were randomly assigned to cue-types (e.g., A, B, C. . .K) for each participant. There was no time limit to make a prediction. Once a prediction had been made, participants indicated their confidence using the confidence scale (see Measures).

The scrollbar was then inactivated, and corrective feedback was provided onscreen for 1.5 s. Specifically, participants were shown Mrs. X's actual antibody level alongside their own estimate on the visual analog scale (they were not given a numeric value as feedback). The position of the feedback indicator on the visual analog scale was jittered around the middle values of each category. This meant that the value given as feedback was not identical on each +, − or ++ trial: antibody levels were randomly selected from a uniform distribution between 0.4 and 1.3 on each – trial, 2.5 and 3.4 on + trials, and between 4.6 and 5.5 on ++ trials. This meant there was always some degree of uncertainty (and therefore potentially prediction error to drive learning) on each trial.

The order of the trials in each of the three phases was randomized with the constraint that all trial types were shown once before any trial type was shown a second time. There were eight instances of each trial type in each phase, yielding 176 trials in total, and the interval between trials was 0.5 s. The transition between phases was not signaled.

Upon completing phase 3, participants were tested. Participants first completed two forced choice test items (see Measures), between cues A and C, and between cues B and D. The order of presentation of these items was randomized for each individual. They then completed test ratings for each cue A–K (see Measures). The cues were presented individually, and the order was randomized for each individual.

Results

Exclusion Criterion

We first examined whether participants learned the initial training contingencies shown in Phase 1. Outcome predictions on the 0.0 to 6.0 scale were coded in 0.1 increments to yield a score of 0–60 for each trial or test item. Participants' mean outcome predictions for cues A–D in the last half (four trials) of Phase 1 training were averaged, yielding a value between 0 and 60. All of these trials were consistently paired with a serious allergic reaction in Phase 1 (4.0 or above). Any participant with a mean rating for these cues of less than the midpoint of the response scale (i.e., less than 30 out of 60) was excluded. This resulted in the removal of nine participants (13%). The remaining analyses were performed on the data from the remaining 59 participants. It is worth noting that the removal of these participants from the statistical analysis did not change the pattern of means in any inferential test in either Experiment 1 or 2. Instead their removal reduced variance (likely noise) from the data. All inferential statistics controlled the two-tailed Type I error rate at 5%, and confidence intervals were constructed at the same confidence level.

Outcome Prediction Accuracy and Confidence Ratings

Outcome predictions and confidence ratings for the critical cues (A–D) across all three training phases are shown in **Figures 2A,B**. Inspection of the figures indicates that participants rapidly learned the contingencies across each training phase. This was evident in their increasing accuracy and confidence across each training phase. Notably, confidence dropped on the second trial of Phase 2, after participants experienced direct disconfirmation of their prior expectations regarding cue A on the initial trial of Phase 2. However, the question of primary theoretical interest in these data is their initial responses to the AB and CD compounds

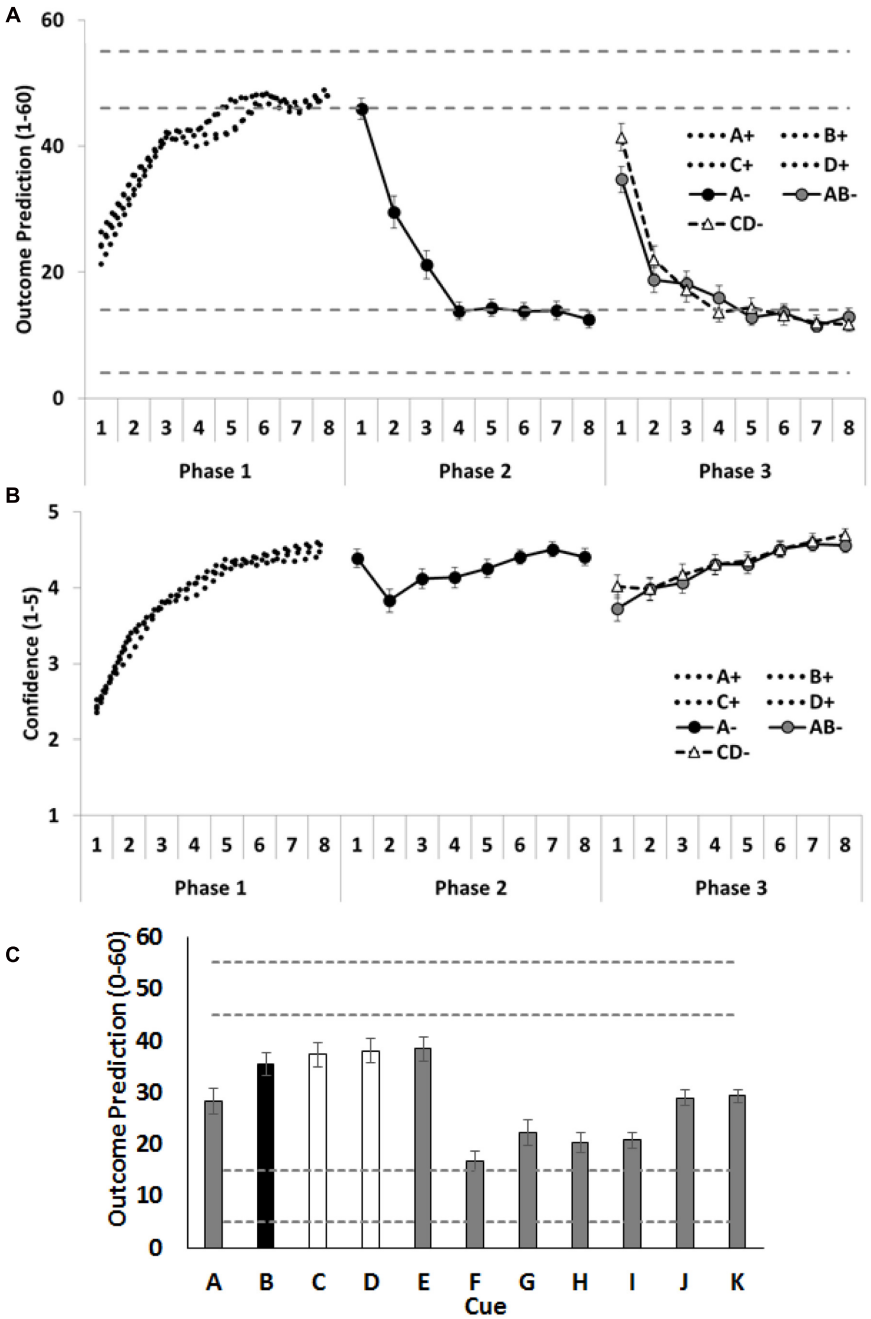


FIGURE 2 | Performance in Experiment 1. (A) Depicts outcome prediction responses across the three training phases for the critical cues A–D. The broken horizontal lines on **(A,C)** indicate the range of outcome values (allergic response severity, 0–60; corresponds to 10 times the 0.0–6.0 value seen in **Figure 1**) that could occur on a ++ trial (the upper range) or on a – trial (the lower range). **(B)** Depicts confidence ratings for these same trial-types across the three training phases. **(C)** Shows people's test ratings for the individual cue test items. In panel C the critical cue (B) is shown as a black column, and the comparison control cues (C and D) are shown as white columns. Error bars indicate standard error of the mean in all panels.

in Phase 3. We hypothesized that people would anticipate the outcome less strongly on the initial AB trial (with one allergenic cue and one extinguished cue) than on the initial CD (with two allergenic cues) trial. This result would be indexed by lower outcome predictions and lower confidence for AB than for CD on the first trial of Phase 3. The initial outcome predictions for compound CD did in fact significantly exceed that for compound AB, $F(1,58) = 5.61$, $p = 0.02$, $\eta_p^2 = 0.09$, CI [1.04, 12.35], but no difference was found between AB and CD on the confidence ratings, $F(1,58) = 3.22$, $p = 0.08$, $\eta_p^2 = 0.05$, CI [−0.03, 0.61].

Test Ratings and Forced Choice Responses

Mean causal ratings for cues A–K are shown in **Figure 2C**. Two orthogonal contrasts were used to examine the amount of extinction for the critical cues A–D. The first contrast compared test ratings for cue B (extinguished in compound with the already extinguished A), with the average of cues C and D (both of which were allergenic when combined into the CD− compound). No significant differences were observed, $F < 1$, $p = 0.41$, $\eta_p^2 = 0.01$, CI [−14.71, 6.13]. Given this absence of a significant difference, the power analysis (Faul et al., 2007, 2009) showed there was sufficient power ($1 - \beta = 0.8$) to detect a small to medium effect size ($f = 0.24$). The implied population effect size for the contrast testing B vs the average of C and D was very small ($f = 0.11$; $\eta_p^2 = 0.01$) and would have required 281 people to find any effect of this magnitude (with $1 - \beta = 0.8$). The second contrast examined whether the additional extinction training given to cue A resulted in more extinction for that cue, than to the less frequently extinguished cues B, C and D; it did, $F(1,58) = 9.18$, $p = 0.004$, $\eta_p^2 = 0.14$, CI [8.95, 43.77].

There were two additional contrasts. The first examined whether participants used the most recent status of each cue or had aggregated over their prior experiences with that cue to generate choice on test. To assess these alternatives, we compared the test ratings given to cue J, which was paired with a minor reaction across all 3 phases, against the average of cues G and H, which were paired with no reaction in Phases 1 and 2, but were paired with severe reaction in the final phase. If people were aggregating their experience across all three phases, they should rate J higher than the recently reinforced G and H; in fact, this was the case, $F(1,58) = 14.47$, $p < 0.001$, $\eta_p^2 = 0.20$, CI [7.35, 23.67]. The second contrast compared two cues, I and K, each of which had been associated with minor reaction, when last shown. However, the prior training was that I had been associated with no reaction, whereas K had no prior training. If participants were influenced by the history of a cue prior to its most recent presentation, they should rate K higher than I; in fact, this was the case, $F(1,58) = 25.20$, $p < 0.001$, $\eta_p^2 = 0.30$, CI [5.09, 11.85]. Taken together, therefore, these results show that participants were influenced by the history of the cue in judging its effectiveness on test, although this does not preclude them from also using the most recent status of a cue in these judgments.

The forced choice data showed a similar pattern. When required to choose between B and D, 32 (54%) participants chose B as the safer food. A binomial test revealed that this did not significantly differ from chance, $p = 0.60$, CI [24.31, 39.69]. By contrast, significantly more participants chose A (68%) as safer than C, $p = 0.009$, CI [23.31, 47.69], indicating that the additional extinction training for A resulted in additional learning for this cue. It is possible that the overall lack of a difference between B and D was obscured by a number of people (at least 32% of the sample) who did not learn that A (extinguished in both Phases 2 and 3) was safer than C (just extinguished in Phase 3). If these participants effectively treated A and C as equivalently extinguished, then there would be no reason to expect a difference in the amounts learned about their partner cues, B and D, respectively. Therefore, we conducted a second analysis of the B versus D forced choice data on only those participants who chose A as safer than C. Of the 44 participants who chose A as safer than C, 28 chose B, the partner of A, as safer than D, the partner of C. This difference was not statistically significant, $p = 0.16$, 95% CI [18.60, 31.40], confirming that the pattern of responding to the target cues, B and D, did not vary with differences in responding to their within-compound partners, A and C.

Discussion

The compound of two allergenic cues (CD−) elicited higher outcome predictions at the beginning of Phase 3, indicating that people initially expected the outcome more on these trials than on the initial AB− trials of Phase 3. This demonstrates that the manipulation was effective and that prediction error was greater across the CD− than the AB− trials; hence, more associative change should have accrued to C and D than to B. However, on the subsequent test, participants did not rate C and D as less allergenic than B nor did they choose D as safer than B. In fact, the direction of the means was in the opposite direction (B > D), both when considering all participants and just those individuals who demonstrated knowledge of A's additional extinction training (by choosing A as safer than C on test). This pattern of results is broadly consistent with previous examinations of compound extinction in human causal learning tasks (Griffiths and Westbrook, 2012; Holmes et al., 2014). It is also consistent with the results from the two aversive conditioning with humans (Lovibond et al., 2000; Vervliet et al., 2007) that also failed to detect any facilitatory effect of extinguishing a compound composed of two aversively conditioned stimuli, as measured by skin conductance and expectancy ratings. This absence of a difference between the target cues (B and D) may be due, in part, to people aggregating over their entire experience with these cues, rather than prioritizing their recent experience (despite the explicit onscreen instructions to do so). Discussion of this issue is withheld to the Section "*General Discussion*."

EXPERIMENT 2

In contrast to the results reported by Lovibond et al. (2000), Vervliet et al. (2007), and Culver et al. (2015) found enhanced extinction for cues trained in compound over cues trained in isolation (on skin conductance and responsiveness to a reinstating outcome, but not on outcome expectancy measures). Culver et al argued that their results were due to having subjected each of the critical cues to extinction before the compound extinction.

Accordingly, our second experiment used a design analogous to that of Culver et al to provide a further examination of the role played by prediction error in extinction of a cue-outcome contingency. This design again involved manipulating prediction error during extinction learning by presenting some cues in compound and others in isolation (see Hendry, 1982). However, in this experiment, the manipulation occured *after* all of the target cues (A−, B−, C−, and D−) had been individually extinguished. Two of those cues (A and B) were then given further extinction in compound (i.e., AB− trials). The rational was that A and B have each retained some association with the outcome, but one that is not sufficient to drive responding on its own. By presenting these two individually ineffectual cues together, their combined capacity to predict the outcome should cross the threshold to elicit renewed prediction of the outcome (Rescorla, 2006). Because these AB− trials therefore elicit some degree of prediction error, this error will drive further extinction learning for these cues. This was tested by comparing the cues given additional compound extinction (A and B), with two control cues (C and D) given the same amount of extinction training but in an individual format (i.e., on separate C− and D− trials). If extinction of causal judgements is regulated by prediction error, extinguished cues that receive additional extinction in compound (A and B) should be treated as safer at test than extinguished cues that received additional extinction in isolation (C and D). As far as we are aware, this hypothesis has not yet been investigated in a human causal learning task.

Materials and Methods

Participants

Seventy six second-year psychology students participated in partial fulfillment of course requirements. The mean age was 20.30 years ($SD = 3.65$), and 60 were female.

Design

The design of the experiment is summarized in the second row of **Table 1**. The four cues (A, B, C, and D) of major interest were each paired with a serious allergic reaction (antibody scores > 4.0) in Phase 1. Then in Phase 2, each of the four cues (A–D) no longer produced that allergic reaction, and were instead followed by no allergic reaction (i.e., normal antibody scores, <20). In the final training phase, Phase 3, A and B were shown together and produced no allergic reaction (AB− trials). The other critical cues, C and D, were each shown individually, and continued to produce no allergic reaction (C− and D− trials).

Measures

The same measures were used as were used in Experiment 1.

Procedure

The procedure was identical to Experiment 1, and only differed with regards to the training contingencies detailed in **Table 1**.

Results

Exclusion Criterion

The same exclusion criterion as used in Experiment 1 was applied to the present data set. It resulted in the removal of data from three participants (4%).

Outcome Prediction Accuracy and Confidence Ratings

As shown in **Figures 3A,B**, Participants rapidly learned the training contingencies: outcome predictions increased (Phase 1) and then decreased (Phases 2 and 3); and confidence in predictions increased across each training phase. Again, our primary theoretical interest concerns how participants treat the critical cues A, B, C and D at the beginning of Phase 3. As predicted, combining the extinguished A and the extinguished B into a compound restored responding, as indicated by the higher outcome predictions for compound AB than for the individually presented C and D. To test this, the average of the outcome predictions on C− and D− trials was compared with the outcome predictions given on AB− trials. Again, the first trial is the data point of most interest, as this is the time at which the outcome predictions based on the compound can be assessed prior to corrective feedback for these predictions. On the first trial, participants gave higher outcome predictions for the AB compound than the average of C− and D− trials, $F(1,72) = 28.78$, $p < 0.001$, $\eta_p^2 = 0.29$, CI [11.74, 25.63]. Moreover, they were less confident about their prediction on the initial AB− trial than on the initial C− and D− trials, $F(1,72) = 10.99$, $p = 0.001$, $\eta_p^2 = 0.13$, CI [0.33, 1.33]. As can be seen in **Figure 3**, this difference between the AB− trials and the C−/D− trials did not persist. By the end of training people were making the same predictions on both the compound (AB−) and the individual (C− and D−) trials with the same levels of confidence.

Test Ratings and Forced Choice Responses

Figure 3C shows the mean causal ratings for cues A–K. A single contrast compared test ratings for the average of the cues, A and B, that had received additional extinction in compound, with the average of the cues, C and D, that had each received additional extinction in isolation. The contrast showed that C and D received significantly lower test ratings, $F(1,72) = 4.58$, $p = 0.04$, $\eta_p^2 = 0.06$, CI [0.42, 11.99] than A and B, indicating that participants learned more about the cues that had been subjected to additional extinction in isolation than in compound. This is the opposite finding to that reported by Leung et al. (2012) using a fear response in rats and Culver et al. (2015) using a skin conductance measure in people.

As in the previous experiment, participants appeared to base their judgements on the aggregated rather than the most recent value of a cue. Specifically, participants rated J, paired throughout with a minor reaction, as more allergenic than G and H, each paired with a severe reaction but only in the final phase, $F(1,72) = 58.59$, $p < 0.001$, $\eta_p^2 = 0.24$, CI [20.17, 34.37]. Participants also rated I, initially paired with no reaction and then with a reaction in Phase 2, as less allergenic than K, paired with a

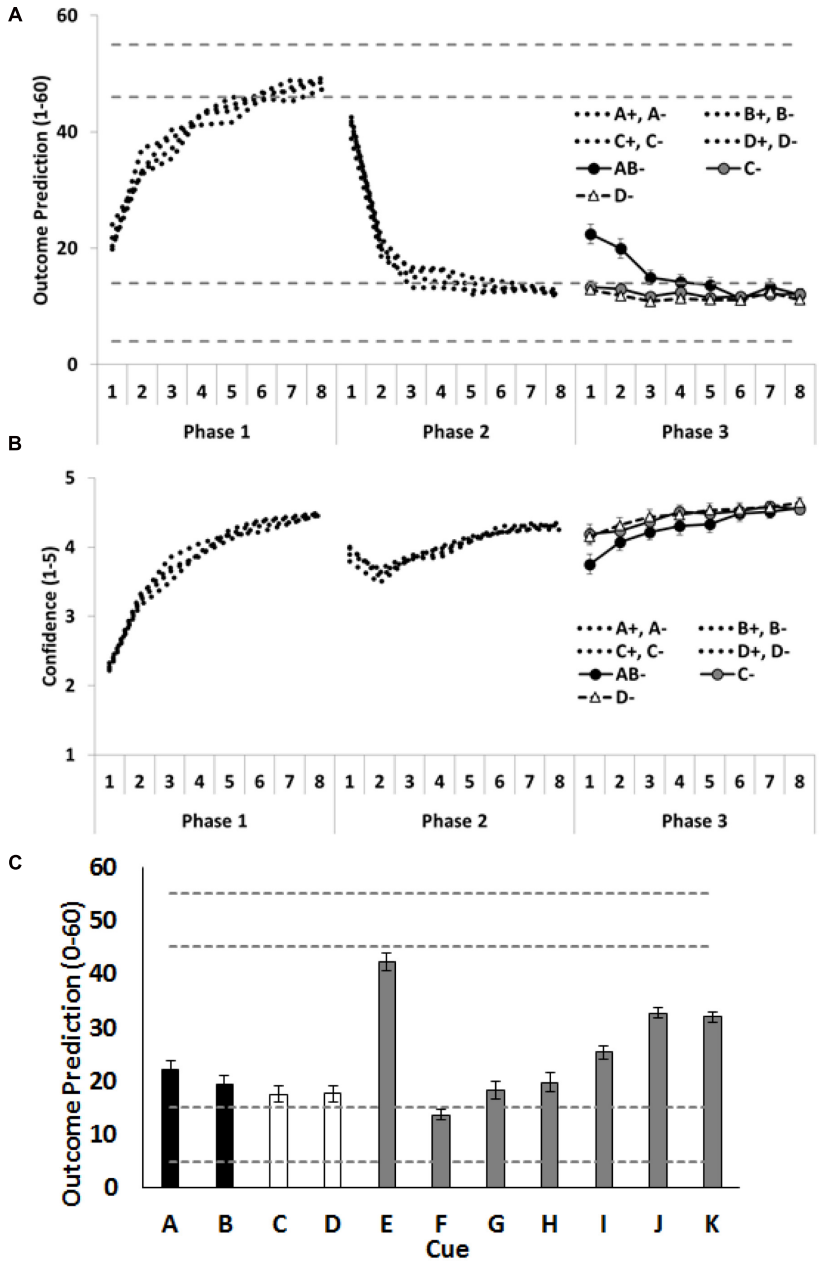


FIGURE 3 | Performance in Experiment 2. (A) Depicts outcome prediction responses across the three training phases for the critical cues A–D. The broken horizontal lines on **(A,C)** indicate the range of outcome values (allergic response severity, 0–60; corresponds to 10 times the 0.0–6.0 value seen in **Figure 1**) that could occur on a ++ trial (the upper range) or on a – trial (the lower range). **(B)** Depicts confidence ratings for these same trial-types across the three training phases. **(C)** People's test ratings for the individual cue test items. The critical cues (A and B) are shown as black columns, and the comparison control cues (C and D) are shown as white columns. Error bars indicate standard error of the mean in all panels.

reaction just in Phase 2, $F(1,72) = 23.20$, $p < 0.001$, $\eta_p^2 = 0.45$, CI [3.87, 9.36].

The forced choice data showed that 35 people chose C as safer than A (52%), and 43 (59%) chose D as safer than B. Because A and B were treated identically, as were cues C and D, inferential statistics were conducted on the choices of A + B versus the choices of C + D. There were more choices of C + D (55%) than of A + B, but this difference was not statistically significant, $p = 0.22$, CI [69.05, 92.94]. We also examined only those participants who chose both A and B (19 people) as compared with those who chose both C and D (27 people; 59%). This difference was also not significant, $p = 0.30$, CI [20.17, 33.83].

Discussion

Compounding two previously extinguished cues (A− and B−) transiently restored outcome predictions. These predictions were significantly higher on the initial AB− trial of Phase 3 than for on the initial C− and D− trials. This means that prediction error across the additional AB− trials should have also been greater than across the additional C− and D− trials and, hence extinction learning about A and B should have been enhanced relative to that learning about C and D. However, this did not occur: in fact, the test measure of outcome expectancy revealed that the individually extinguished C and D were rated as significantly safer (less allergenic) than the otherwise matched, but compound extinguished, A and B. The forced choice items were in the same direction as their outcome expectancy ratings, but no significant differences were observed. In sum, the compound manipulation used to restore responding was successful but the deepening of extinction learning across additional extinction of that compound was not confirmed: if anything, that additional extinction of the compound appeared to impair extinction learning in this human learning analog.

GENERAL DISCUSSION

Two experiments examined whether extinction of cue-outcome contingency knowledge is regulated by an error-correction process: specifically, whether manipulations that maintain or restore outcome expectancies in extinction can facilitate or deepen the learning that occurs when a cue is presented in the absence of its expected outcome. This deepening has been observed in extinction of conditioned fear in rats (Leung et al., 2012) and extinction of skin conductance responses in people (Culver et al., 2015). This effect is predicted by theories of associative learning (e.g., Rescorla and Wagner, 1972; Wagner, 1981) which hold that all the cues present on a trial are used to calculate the error whose size determines the amount of associative change and whose sign (positive or negative) determines the nature of the change (excitatory in the case of acquisition or inhibitory in the case of extinction).

In each of the two experiments, we used a different manipulation to maintain or restore outcome expectancies across extinction of the cue-outcome contingencies. In Experiment 1, a target cue, D, was extinguished in compound with a non-extinguished cue, C, and the consequences for its extinction were assessed relative to a control cue, B, extinguished in compound with an already-extinguished cue, A. Critically, this manipulation was effective in generating differences in responding such that the CD compound was treated as more allergenic than the AB compound, which should have served to increase the size of the prediction error on CD− trials relative to AB− trials. However, the levels of test responding to B and D revealed no evidence that the larger error on CD− trials had deepened the extinction of D relative to that of its control cue, B: both cues were rated as equally allergenic, and when forced to make a choice, equal numbers of people chose B as more allergenic than D, and D as more allergenic than B. Thus, just as Lovibond et al. (2000) and Vervliet et al. (2007) failed to find any evidence for facilitated extinction of skin conductance responses or expectancy ratings to a cue predictive of shock, Experiment 1 failed to find any evidence for facilitated extinction of cue-outcome contingency knowledge in a causal judgment task.

These results clearly offer no support for the hypothesis that extinction of cue-outcome contingency knowledge is regulated by prediction error. However, they should not be taken as evidence *against* that hypothesis. The design used in Experiment 1, which is based on that used by Lovibond et al. (2000) and Vervliet et al. (2007), is one for which the predictions of error-correction theories are parameter dependent. Specifically, as the target cue, D, was only ever extinguished in compound with a non-extinguished partner, C, error-correction theories predict that its extinction should have been facilitated (i.e., participants should have abandoned responding to D at a faster rate than they abandoned responding to B), but, critically, that extinction of D would not necessarily have been deepened: that is, such theories hold that with sufficient extinction the net strengths of D and B at the end of extinction will in fact be equal. Hence, rather than showing that extinction of cue-outcome contingency knowledge is *not* regulated by prediction error, an alternative explanation for the results of Experiment 1 is that B and D had been extinguished to their common low asymptote, and hence, there was no opportunity for detecting any facilitation of extinction to D relative to B. However, it is noteworthy that most other cues were given lower ratings at test than either B or D (see **Figure 1**), which diminishes the conclusion that these cues were both at their lowest, asymptotic value.

In any case, there is no ambiguity in error-correction theories' predictions of deepened extinction in Experiment 2; a deepening that has been found with affective reactions in aversive conditioning procedures with rats (Leung et al., 2012) and people (Culver et al., 2015). In this experiment, four allergenic cues, A, B, C, and D, were each presented alone during an initial phase of extinction. The target cue, B, then received additional extinction in compound with one of the other extinguished cues, A, while control cues C and D continued to be extinguished alone. Critically, the compounding of two already-extinguished cues, AB, restored the expectation of the outcome relative to continued presentations of C and D alone: that is, the AB compound was treated as more allergenic than presentations of either C or D alone, and, hence, the size of the error on the AB− trials should have been greater than on C− and D− trials. However, here again, ratings of the individual cues at test revealed no evidence that the larger error on AB− trials had deepened the extinction of A and B relative to that of the control cues, C and D. In fact, if anything, we observed the opposite result: A and B were rated as more allergenic than C and D. Thus, unlike the findings reported by Leung et al. (2012) and Culver et al. (2015), the present experiment failed to find any evidence for deepened extinction of (affectively neutral) cue-outcome contingency knowledge in a causal judgment task.

One way of reconciling the findings reported by Leung et al. (2012) and Culver et al. (2015) with those reported in the present study is to assume that there are differences across the protocols (aversive conditioning versus causal judgments) in the extent to which the effects of compound extinction

generalize to testing (e.g., Pearce, 1987). Specifically, there was less generalization of compound extinction in the present study than in the two previous ones, possibly as a function of differences in cue duration and trial rate, and/or the types of association formed in extinction (affective versus contingency knowledge). For example, in the Leung et al. (2012) and Culver et al. (2015) studies, the cues were of fixed duration (30 and 8 s, respectively) and the interval between trials in acquisition and extinction was relatively long (120 s and ∼25 s, respectively); whereas in the present study (and other studies of human causal judgments), cues were presented on screen for as long as it took participants to respond (typically, 1–2 s), and the interval between the response and the subsequent trial was much shorter (0.5 s). It has previously been shown that both of these parameters can influence the likelihood of inhibitory or excitatory learning in procedures where both types of learning are possible (i.e., second-order conditioning; Karazinov and Boakes, 2007); perhaps this may also influence the propensity to generalize from configural to elemental representations.

Another way of expressing the same point is that the methods of testing used here were not sufficiently sensitive to detect the effects of compound extinction reported previously. Indeed, the self-rated test items in this task have an inherent limitation with respect to the information participants are likely to use when answering them. Collins and Shanks (2002) noted that when people are asked to rate the likelihood that an outcome will follow a cue, their answer critically depends on when they are asked. If asked during the training phase, people are more influenced by their recent experience with the cue and the outcome, whereas if asked at the end of training (in a test phase), people are more likely to aggregate across all of their experiences with the cue and outcome. Such aggregations would minimize any differences in recent extinction training, such as those investigated here. To minimize the likelihood of people responding at test based on averaging, we emphasized to participants that they should rely upon their recent experience (how would Mrs. X react *now* if she ate this food). This was achieved by adjusting the cover story of the allergist task, and altering the wording of the test question and response items. First, people were told from the outset that Mrs. X would soon undertake a medical procedure during which time she could not afford to have an allergic reaction. Therefore, people were asked to review her recent meal intake and allergic responses (the training phase), before acting as allergists to advise which foods were most likely to be safe for her during the procedure (the test phase). Each trial presented an incrementally increasing date on the screen, and the date of the test phase items followed immediately those of the training phase. Second, the wording of the instructions for the test phase again emphasized people needed to indicate which foods were safe for Mrs. X "right now." The test question shown on each test item asked "How likely is this food to produce an allergic reaction in Mrs. X right now?" and the anchors on the response scrollbar similarly included the word now.

Despite these efforts, the analyses of the distractor cues (G, H, I, J, K) in both experiments suggest that the test ratings were influenced by their experience with each cue prior to the final phase in which that cue was shown. For instance, at the time of test, cues G and H were paired most recently with a strong allergic reaction (outcome ++) where cues J and K were most recently paired with a mild allergic reaction (outcome +). Yet across both experiments, cue J was rated higher than G and H at test. Such data suggest that our prominent, repeated verbal instructions were not, or not completely, successful in directing participants to base their test ratings just on their recent experience with a cue rather than on their history of experience with that cue.

CONCLUSION

The present study showed that extinction of a target cue in compound with either a second allergenic cue (Experiment 1) or a second extinguished cue (Experiment 2) led to a maintenance or restoration of outcome expectancies across compound extinction. However, even though both manipulations increased the prediction error during the critical phase of compound extinction, neither facilitated nor deepened extinction learning of cue-outcome contingency knowledge. These results are similar to those reported by Holmes et al. (2014), and on the face of it, stand in contrast to findings reported by Leung et al. (2012) and Culver et al. (2015) showing that extinction of affective reactions to a target cue can be deepened. It is possible that this difference in conclusion relies upon the parameters of the acquisition, extinction and test procedures used, and also upon people's propensity to use all of their prior experience with a cue, rather than only their most recent experiences. If so the efficacy of enhancing exposure therapy using these methods may depend critically on the specific spacing, duration and format of both the exposure sessions and any anxiety-relevant events that have occurred in the past. Because a number of these properties are typically outside of the therapists' control, it remains unclear whether the present prediction-error enhancing methods will readily generalize to clinical practice. These questions remain for future research.

AUTHOR CONTRIBUTIONS

OG contributed to: the construction of materials, the design of the experiment, the statistical analysis, the interpretation of the data, the preparation of the manuscript. NH contributed to: the statistical analysis, the interpretation of the data, the preparation of the manuscript. RW contributed to: the design of the experiment, the interpretation of the data, the preparation of the manuscript.

ACKNOWLEDGMENTS

This work was supported by an Australian Research Council (ARC) Discovery Early Career Research Award (DE150100667) awarded to OG, and ARC Discovery (D123456789) grant awarded to RW and NH.

REFERENCES

Aitken, M. R. F., Larkin, M. J. W., and Dickinson, A. (2000). Super-learning of causal judgements. *Q. J. Exp. Psychol.* 53B, 59–81. doi: 10.1080/027249900392995

Chapman, G. B., and Robbins, S. J. (1990). Cue interactions in human contingency judgment. *Mem. Cogn.* 18, 537–545. doi: 10.3758/BF03198486

Clark, D. M. (1986). A cognitive approach to panic. *Behav. Res. Ther.* 24, 461–470. doi: 10.1016/0005-7967(86)90011-2

Collins, D. J., and Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Mem. Cogn.* 30, 1138–1147. doi: 10.3758/BF03194331

Culver, N. C., Vervliet, B., and Craske, M. G. (2015). Compound extinction using the Rescorla–Wagner model to maximize exposure therapy effects for Anxiety disorders. *Clin. Psychol. Sci.* 3, 335–348. doi: 10.1177/2167702614542103

Dickinson, A., Shanks, D., and Evenden, J. (1984). Judgement of act-outcome contingency: the role of selective attribution. *Q. J. Exp. Psychol.* 36, 29–50. doi: 10.1080/14640748408401502

Eddy, K. T., Dutra, L., Bradley, R., and Westen, D. (2004). A multidimensional meta-analysis of psychotherapy and pharmacotherapy for obsessive-compulsive disorder. *Clin. Psychol. Rev.* 24, 1011–1030. doi: 10.1016/j.cpr.2004.08.004

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Feske, U., and Chambless, D. L. (1995). Cognitive behavioral versus exposure only treatment for social phobia: a meta-analysis. *Behav. Ther.* 26, 695–720. doi: 10.1016/S0005-7894(05)80040-1

Griffiths, O., Johnson, A., and Mitchell, C. J. (2011). Negative transfer in human associative learning. *Psychol. Sci.* 22, 1198–1204. doi: 10.1177/0956797611419305

Griffiths, O., Mayrhofer, R., Nagel, J., and Waldmann, M. R. (2009). "Causal schema-based inductive reasoning," in *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, Amsterdam, 691–696.

Griffiths, O., and Mitchell, C. J. (2009). Selective attention in human associative learning and recognition memory. *J. Exp. Psychol. Gen.* 137, 626–648. doi: 10.1037/a0013685

Griffiths, O., and Westbrook, R. F. (2012). A common error term regulates acquisition by not extinction of causal learning. *Learn. Behav.* 40, 207–221. doi: 10.3758/s13420-011-0056-0

Hendry, J. S. (1982). Summation of undetected excitation following extinction of the CER. *Anim. Learn. Behav.* 10, 476–482. doi: 10.3758/BF03212287

Holmes, N., Griffiths, O., and Westbrook, R. F. (2014). Role of prediction error in extinction of causal judgments in people. *Learn. Behav.* 42, 289–303. doi: 10.3758/s13420-014-0146-x

Holmes, N. M., and Westbrook, R. F. (2013). Extinction of reinstated or ABC renewed fear responses renders them resistant to subsequent ABA renewal. *J. Exp. Psychol. Anim. Behav. Process.* 39, 208–220. doi: 10.1037/a0031986

Kamin, L. J. (1969). "Selective association and conditioning," in *Fundamental Issues in Associative Learning*, eds N. J. Mackintosh and W. K. Honig (Halifax, NS: Dalhousie University Press), 42–64.

Karazinov, D. M., and Boakes, R. A. (2007). Second-order conditioning in human predictive judgements when there is little time to think. *Q. J. Exp. Psychol.* 60, 448–460. doi: 10.1080/17470210601002488

Leung, H. T., Reeks, L. M., and Westbrook, R. F. (2012). Two ways to deepen extinction and the difference between them. *J. Exp. Psychol. Anim. Behav. Process.* 38, 394–406. doi: 10.1037/a0030201

Leung, H. T., and Westbrook, R. F. (2008). Spontaneous recovery of extinguished fear responses deepens their extinction: a role for error-correction mechanisms. *J. Exp. Psychol. Anim. Behav. Process.* 34:461. doi: 10.1037/0097-7403.34.4.461

Leung, H. T., and Westbrook, R. F. (2010). Increased spontaneous recovery with increases in conditioned stimulus alone exposures. *J. Exp. Psychol. Anim. Behav. Process.* 36, 354–367. doi: 10.1037/a0017882

Lovibond, P. F. (2004). Cognitive processes in extinction. *Learn. Mem.* 11, 495–500. doi: 10.1101/lm.79604

Lovibond, P. F., Davis, N. R., and O'Flaherty, A. S. (2000). Protection from extinction in human fear conditioning. *Behav. Res. Ther.* 38, 967–983. doi: 10.1016/S0005-7967(99)00121-7

McAndrew, A., Jones, F. W., McLaren, R. P., and McLaren, I. P. L. (2012). Dissociating expectancy of shock and changes in skin conductance: an investigation of the perruchet effect using an electrodermal paradigm. *J. Exp. Psychol. Anim. Behav. Process.* 38, 203–208. doi: 10.1037/a0026718

McConnell, B. L., Miguez, G., and Miller, R. R. (2013). Extinction with multiple excitors. *Learn. Behav.* 41, 119–137. doi: 10.3758/s13420-012-0090-6

Mitchell, C. J., De Houwer, J., and Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behav. Brain Sci.* 32, 183–198. doi: 10.1017/S0140525X09000855

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychol. Rev.* 94:61. doi: 10.1037/0033-295X.94.1.61

Pearce, J. M., and Wilson, P. N. (1991). Effects of extinction with a compound conditioned stimulus. *J. Exp. Psychol. Anim. Behav. Process.* 17, 151.

Raes, A. K., De Houwer, J., Verschuere, B., and De Raedt, R. (2011). Return of fear after retrospective inferences about the absence of an unconditioned stimulus during extinction. *Behav. Res. Ther.* 49, 212–218. doi: 10.1016/j.brat.2010.12.004

Reberg, D. (1972). Compound tests for excitation in early acquisition and after prolonged extinction of conditioned suppression. *Learn. Motiv.* 3, 246–258. doi: 10.1016/0023-9690(72)90021-5

Rescorla, R. A. (1969). "Conditioned inhibition of fear," in *Fundamental Issues in Associative Learning*, eds N. J. Mackintosh and W. K. Honig (Halifax, NS: Dalhousie University Press.), 65–89.

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. *Learn. Motiv.* 1, 372–381. doi: 10.1016/0023-9690(70)90101-3

Rescorla, R. A. (2006). Deepened extinction from compound stimulus presentation. *J. Exp. Psychol. Anim. Behav. Process.* 32, 135.

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Salkovskis, P. M. (1991). The importance of behaviour in the maintenance of anxiety and panic: a cognitive account. *Behav. Psychother.* 19, 6–19. doi: 10.1017/S0141347300011472

Taylor, S. (1996). Meta-analysis of cognitive-behavioral treatments for social phobia. *J. Behav. Ther. Exp. Psychiatry* 27, 1–9. doi: 10.1016/0005-7916(95)00058-5

Thomas, B. L., and Ayres, J. J. (2004). Use of the ABA fear renewal paradigm to assess the effects of extinction with co-present fear inhibitors or excitors: implications for theories of extinction and for treating human fears and phobias. *Learn. Motiv.* 35, 22–52. doi: 10.1016/S0023-9690(03)00040-7

Vervliet, B., Vansteenwegen, D., Hermans, D., and Eelen, P. (2007). Concurrent excitors limit the extinction of conditioned fear in humans. *Behav. Res. Ther.* 45, 375–383. doi: 10.1016/j.brat.2006.01.009

Wagner, A. R. (1969). "Stimulus validity and stimulus selection in associative learning," in *Fundamental Issues in Associative Learning*, eds N. J. Mackintosh and W. K. Honig (Halifax, NS: Dalhousie University Press), 90–122.

Wagner, A. R. (1981). SOP: a model of automatic memory processing in animal behavior. *Inform. Process. Anim. Mem. Mech.* 85, 5–47.

Witnauer, J. E., and Miller, R. R. (2012). Associative status of the training context determines the effectiveness of compound extinction. *J. Exp. Psychol. Anim. Behav. Process.* 38, 52–65. doi: 10.1037/a0026333

Zeng, Q., Jia, Y., Wang, Y., Zhang, J., Liu, C., and Zheng, X. (2015). Retrospective reversal of extinction of conditioned fear by instruction. *Conscious. Cogn.* 35, 171–177. doi: 10.1016/j.concog.2015.05.011

Conflict of Interest Statement: 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.

Copyright © 2017 Griffiths, Holmes and Westbrook. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 June 2017
doi: 10.3389/fpsyg.2017.00926

Learning to Detect Triggers of Airway Symptoms: The Role of Illness Beliefs, Conceptual Categories and Actual Experience with Allergic Symptoms

Thomas Janssens, Eva Caris, Ilse Van Diest and Omer Van den Bergh*

Health Psychology, KU Leuven, Leuven, Belgium

OPEN ACCESS

Edited by:
Anna Thorwart,
Philipps University of Marburg, Germany

Reviewed by:
Gabrielle Weidemann,
Western Sydney University, Australia
Mitchell Rabinowitz,
Fordham University, United States

****Correspondence:***
Thomas Janssens
thomas.janssens@kuleuven.be

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *27 October 2016*
Accepted: *19 May 2017*
Published: *07 June 2017*

Citation:
Janssens T, Caris E, Van Diest I and Van den Bergh O (2017) Learning to Detect Triggers of Airway Symptoms: The Role of Illness Beliefs, Conceptual Categories and Actual Experience with Allergic Symptoms. Front. Psychol. 8:926. doi: 10.3389/fpsyg.2017.00926

Background: In asthma and allergic rhinitis, beliefs about what triggers allergic reactions often do not match objective allergy tests. This may be due to insensitivity for expectancy violations as a result of holding trigger beliefs based on conceptual relationships among triggers. In this laboratory experiment, we aimed to investigate how pre-existing beliefs and conceptual relationships among triggers interact with actual experience when learning differential symptom expectations.

Methods: Healthy participants ($N = 48$) received information that allergic reactions were a result of specific sensitivities versus general allergic vulnerability. Next, they performed a trigger learning task using a differential conditioning paradigm: brief inhalation of CO_2 enriched air was used to induce symptoms, while participants were led to believe that the symptoms came about as a result of inhaled allergens (conditioned stimuli, CS's; CS+ followed by symptoms, CS− not followed by symptoms). CS+ and CS− stimuli either shared (e.g., birds-mammals) or did not share (e.g. birds-fungi) category membership. During Acquisition, participants reported symptom expectancy and symptom intensity for all triggers. During a Test 1 day later, participants rated symptom expectancies for old CS+/CS− triggers, for novel triggers within categories, and for exemplars of novel trigger categories. Data were analyzed using multilevel models.

Findings: Only a subgroup of participants ($n = 22$) showed differences between CO_2 and room air symptoms. In this group of responders, analysis of symptom expectancies during acquisition did not result in significant differential symptom CS+/CS− acquisition. A retention test 1 day later showed differential CS+/CS− symptom expectancies: When CS categories did not share category membership, specific sensitivity beliefs improved retention of CS+/CS− differentiation. However, when CS categories shared category membership, general vulnerability beliefs improved retention of CS+/CS− differentiation. Furthermore, participants showed some selectivity in generalization of symptom expectancies to novel categories, as symptom expectancies did not generalize to

novel categories that were unrelated to CS+ or CS− categories. Generalization to novel categories was not affected by information about general vulnerability or specific sensitivities.

Discussion: Pre-existing vulnerability beliefs and conceptual relationships between trigger categories influence differential symptom expectancies to allergic triggers.

Keywords: asthma triggers, contingency learning, generalization (psychology), expectancy violation, illness perceptions

INTRODUCTION

Asthma and allergic rhinitis are chronic conditions that are characterized by an allergic or hyperreactive response of the airways to a variety of triggers (Bousquet et al., 2012; Global Initiative for Asthma (GINA), 2016). Because treatment for these conditions is currently not available, management strategies are suggested to reduce the manifestation of symptoms and increase clinical control (Global Initiative for Asthma (GINA), 2016). These management strategies are multifaceted, and include pharmacological strategies (often a combination of preventer and reliever medication) as well as behavioral strategies of trigger identification and subsequent avoidance as a way to obtain control over symptoms (Global Initiative for Asthma (GINA), 2016). However, despite these treatment options day-to-day control over symptoms is often poor (Rabe et al., 2004; Peters et al., 2007).

One reason for the lack of day-to-day symptom control may be the difficulties that arise when implementing trigger identification and behavioral avoidance strategies (Janssens and Ritz, 2013). These latter strategies rely on the perception of spatio-temporal contingencies between the presence of triggers and subsequent emergence of asthmatic or allergic symptoms in order to allow prediction of symptoms and accurate avoidance of triggers. In other words, based on medical information and personal experiences, patients construct trigger beliefs to guide their (future) behavior. Interestingly, trigger beliefs often do not match with the results of a structured trigger evaluation procedure, with both false positives and false negatives being observed (Li et al., 2000; Smith et al., 2009). Furthermore, in day-to-day asthma management, individuals with asthma often report being uncertain about their personal triggers and trigger avoidance strategies (Caress et al., 2002; Trollvik and Severinsson, 2004). In addition, individuals show a marked variation in the type and number of asthma triggers they identify, with a higher number of self-identified asthma triggers being associated with worse asthma outcomes, even when controlling for other measures of asthma severity (Ritz et al., 2006, 2016; Janssens and Harver, 2015). Taken together, these findings suggest difficulties and inaccuracies in the process of asthma trigger identification or the detection of trigger-symptom contingencies. Moreover, literature on symptom perception suggests that these beliefs about trigger-symptom contingencies may in turn bias perception of respiratory symptoms (Janssens et al., 2009; von Leupoldt and Dahme, 2012), which may lead to even more difficulties in trigger identification.

Previously, we have highlighted similarities between asthma trigger learning and other contingency learning tasks that occur in a motivational context, such as the identification of danger and safety that occurs within the context of fear learning (Janssens and Ritz, 2013; Janssens et al., 2015). Building upon these similarities, we have explored generalization of symptom-trigger contingencies as a potential mechanism of the observed inaccuracies in asthma trigger identification. Similar to conceptualization of generalization in the context of anxiety and fear, generalization of trigger beliefs may serve an adaptive purpose in that it helps to transfer knowledge that is gained from experience to similar instances which have not (yet) been experienced, therefore limiting the risk of adverse symptom outcomes. However, generalization may also be considered excessive or maladaptive when innocuous stimuli are treated as threatening, especially if the associated symptoms and behavioral responses interfere with day to day functioning or quality of life (Dunsmoor et al., 2009; van Meurs et al., 2014). An illustrative example in the field of allergy is the avoidance of tree nuts by individuals that show a sensitivity to peanut allergens. This avoidance seems sensible, based on considerable similarities between peanuts and tree nuts. However, a recent review of the available evidence for this strategy shows that avoidance of all tree nuts in individuals with peanut allergy may be overly precautious (Brough et al., 2015).

So far, in associative learning research, most research on generalization has studied perceptual similarities as a basis for generalization. However, recent research has also explored the role of higher order cognitions such as category membership and stimulus typicality as a basis for generalization, showing that participants can use their pre-existing knowledge about categories as a basis for fear generalization (Dunsmoor and Murphy, 2015; Dymond et al., 2015). Based on these developments in fear generalization research, we previously have adapted an associative learning or conditioning paradigm focusing on category based fear learning (Dunsmoor et al., 2012) into a lab method to investigate category-based respiratory trigger learning. Briefly, this method consists of the presentation of pictures, which are unique exemplars of two different allergen categories (e.g., mammals and flowers). Exemplars of one category (conditioned stimuli, CS+) predict onset of respiratory symptoms, whereas exemplars of the other category (CS−) are never followed by symptoms. Using this method, we observed generalization of trigger beliefs to novel category exemplars, as well as to exemplars of categories that were similar of the original trigger categories, providing a proof of concept that trigger beliefs are shaped by pre-existing conceptual knowledge

(Janssens et al., 2015). Moreover, an important finding of this study was that generalization of symptom expectancies to novel CS+ exemplars was increased if participants had experienced CS+ and CS− categories that were more similar (e.g., mammals and birds), compared to categories that were more different (e.g., mammals and molds). We interpreted this finding as an effect of discrimination learning on the inferred relevance of category features as basis for generalization, which is in line with other studies that have showed an impact of either inferred or instructed feature relevance on feature based fear generalization, and support feature-extraction or rule-based accounts of generalization (Vervliet et al., 2010; Vervliet and Geens, 2014; Ahmed and Lovibond, 2015a,b).

The role of category identification and feature extraction in the generalization of cue-outcome contingencies prompts investigation into the potential role of other complex cognitive mechanisms in changing the course of generalization. More specifically, it may provide opportunities to link research on generalization with the large body of research on the role of illness-related beliefs in the context of symptom perception and disease-related behaviors (Leventhal et al., 1980; Hagger and Orbell, 2003). In asthma, research within this framework has been successful in highlighting the role of beliefs about symptom chronicity, controllability, and medication necessity and concerns, in explaining individual differences in symptom perception and medication use patterns (Horne and Weinman, 1999; Halm et al., 2006; Kaptein et al., 2010). However, research into beliefs about causality and beliefs about trigger-symptom causal chains have been limited. An exception to this is a study by McQuaid et al. (2002), who studied the cognitive complexity of causal understanding in children with asthma and their parents. In this study, participants were asked to elaborate on the question "what causes your asthma," and "how does this trigger cause asthma symptoms." Results of this study showed a variety of complexity of responses, ranging from phenomism (no differentiation between cause and effect) to complex psychophysiological causal models, with more complex understanding of causal chains in asthma being associated with better treatment strategies.

Building upon this study, the aim of our research was to investigate the relationship between beliefs about causality in asthma and the way individuals integrate real life experiences into models of symptom-trigger contingency. Our study provides a lab based analog for a common task in the initial treatment phase of allergy management: individuals receive information about what asthma is, and are confronted with a variety of potential triggers, that are linked to adverse outcomes (airway symptoms) in a probabilistic way. In line with our focus on causality, we chose to focus on beliefs that link asthma triggers to a general vulnerability vs. beliefs that focus on asthma triggers as very specific indicators of specific airway sensitivities, thereby mimicking different information that may be given to patients with allergic conditions by their physician or information individuals may find on the internet (Smith et al., 1998; Croft and Peterson, 2002; Huckvale et al., 2012). The actual contingencies that were presented in the task did not fully confirm or disconfirm this prior information, in that during acquisition, each potential trigger that was presented was unique. However, participants could use their knowledge of category membership and category relations to infer differences in trigger-symptom contingencies at a category level. We hypothesized that a focus on general vulnerability would hinder differentiation between triggers and non-triggers, whereas a focus on specific sensitivities would improve differentiation between triggers and non-triggers. Furthermore, in line with our previous findings on differential acquisition of trigger beliefs, we expected that the use of CS− trigger categories that were more similar to the CS+ categories would enhance differentiation between CS+ and CS− trigger beliefs.

MATERIALS AND METHODS

Participants

The study was approved by the Social and Societal Ethics Committee at KU Leuven and the Ethical Review Board of Leuven University Hospitals (study ID: ML10101). Participants were 48 healthy volunteers (15 male, aged 17–38), recruited from the student population. Psychology students received course credit for participation in the experiment. The other participants received 12 euros.

Exclusion criteria were self-reported allergies, hay fever, asthma or other lung disease, heart disease, epilepsy, other severe medical or psychiatric illnesses and the presence of electronic implants. Furthermore, participants were excluded if their lung function (forced expiratory volume in 1 s) was below 80% of their predicted value.

Materials

Measures

Symptom expectancy was measured using a visual analog scale (VAS) anchored at *definitely no symptoms* and *definite symptoms*. For symptom intensity and unpleasantness, VAS were used with the anchors *not at all intense/unpleasant* and *maximal imaginable intensity/unpleasantness*.

During the online retention/generalization test, for all pictures in the trigger stimulus set, participants rated whether they had seen the picture during the lab task, or whether it was novel. Furthermore, symptom probabilities were assessed on an 11-point scale ranging from 0% (will not experience symptoms), to 100% (will definitely experience symptoms).

The Positive and Negative Affect Schedule [PANAS; (Watson et al., 1988), Dutch version (Engelen et al., 2006)] was used to assess trait positive affect and trait negative affect. The PANAS is a 20 item scale consisting of positive and negative emotion words. For each of the items, participants indicate on a 5-point scale, ranging from *very little* to *very much,* to which extend they experience each of these feelings in their daily lives.

Suffocation fear was measured using the suffocation scale of the Dutch Claustrophobia Questionnaire (CLQ; Van Diest et al., 2010). This scale consists of 14 situations that may elicit suffocation fears. Participants rate how fearful they would feel in each of the situations, on a 5-point scale ranging from *not at all fearful* to *extremely fearful*.

Stimuli

Asthma trigger stimuli consisted of four categories of potential asthma triggers: mammals, birds, flowers, and molds. This is the same stimulus set that we have used in previous research (Janssens et al., 2015). Each category consists of 20 unique pictures, and stimulus categories can be organized into two hierarchical categories: "animals" (mammals; birds) and "plants"(flowers; molds), creating the potential for constructing acquisition trigger sets with CS's that are conceptually more/less similar. The difference in similarity between category pairs was tested and confirmed in previous research (Janssens et al., 2015). Allocation of CS+/− categories during acquisition was counterbalanced across participants, according to **Table 1**.

Asthma information was embedded in the informed consent form. In the general vulnerability condition, this consisted of information that asthma was an allergic condition, and that allergic responses to allergy triggers were an indication of a general vulnerability making it necessary to avoid all potential asthma triggers. The condition highlighting specific sensitivities consisted of information that asthma was the result of an allergic response to specific allergens, and that careful investigation of triggers and non-triggers was possible, so that individuals with asthma do not need to avoid a variety of potential triggers.

Apparatus

Lung function was measured using a spirometer (Jaeger Masterscope; Hoechberg, Germany) prior to the actual start of the experimental breathing trials. For the latter trials, a valve was used for switching between the regular room air and the CO_2-enriched air. The CO_2-enriched air consisted of a mixture of 7.5% CO_2, 21% O_2, and 71.5% N_2 fed into a meteorological balloon. Short-term inhalation of CO_2-enriched air affects respiration, increasing breathing frequency and volume, and feelings of breathlessness, mimicking aspects of asthma symptoms (De Peuter et al., 2008; Janssens et al., 2011). The participants breathed into a mask connected to the valve through an antibacterial filter. The mask was also connected to a capnograph (Nonin LifeSense, Leek, The Netherlands) and a pneumotachograph (Fleisch No. 2, fg-deutschland; Hechingen, Germany). Affect 4.0 software (Spruyt et al., 2010) was used for stimulus presentation and to record participant responses and capnograph and pneumotachograph signals.

TABLE 1 | Trigger stimulus set and category relations.

CS+ and CS− conceptually similar			
CS+	**CS−**	**Gu**	
Flowers	Molds	Birds, Mammals	
Molds	Flowers	Birds, Mammals	
Birds	Mammals	Flowers, Molds	
Mammals	Birds	Flowers, Molds	
CS+ and CS− conceptually dissimilar			
CS+	**CS−**	**G+**	**G−**
Flowers	Birds	Molds	Mammals
Molds	Mammals	Flowers	Birds
Birds	Molds	Mammals	Flowers
Mammals	Flowers	Birds	Molds

G+, generalization stimuli conceptually related with CS+; G−, generalization stimuli conceptually related with CS; Gu, generalization stimuli unrelated to CS+ or CS−.

Procedure

When participants arrived at the laboratory, they received oral and written information about the experiment. Participants were told that they would inhale a series of aerosols, each containing a mixture of air and a specific allergen, and that there was a risk of the occurrence of respiratory symptoms during these breathing trials. The information about the experiment also included our asthma information manipulation, and participants were randomly assigned to receive information focusing on general vulnerability or specific sensitivities.

After reviewing the information and exclusion criteria, participants completed informed consent. Subsequently, lung function was measured.

Subsequently, trigger acquisition trials started using a similar trial-unique acquisition procedure as in Janssens et al. (2015). The experimenter left the room and participants received 20 breathing trials. Each breathing trial followed the same pattern. First, a novel picture of a potential asthma trigger was shown, indicating to the participant that this allergen would be presented during the breathing trial (although in reality no allergens were present and symptom onset and trigger-symptom contingency was experimentally controlled). Ten pictures randomly chosen from each the CS+ and CS− trigger category was used for this purpose. After presentation of the picture, participants rated symptom expectancy using the VAS expectancy scale. Subsequently, participants were instructed to breathe through the mask, while the picture remained visible. Through the mask, the participants inhaled either regular room air either CO_2-enriched air. For 6 out of the 10 CS+ trials, participants inhaled CO_2-enriched air followed after the pictures. In all other trials, participants inhaled room air. After 60 s, participants could take off the mask, and rated symptom intensity and unpleasantness, using the intensity and unpleasantness VAS scales. Ratings were followed by a 2-min recovery phase, after which participants were prompted to start a new breathing trial.

One day after trigger acquisition, participants filled out an online survey. The survey consisted of the PANAS and the Suffocation scale of the CLQ, as well as recognition and symptom probability ratings of the full trigger picture set. Trigger pictures were presented in random order. After completion of the survey, participants were debriefed.

Data Reduction and Data Analysis

In order to obtain data about breathing behavior, pneumotachograph and capnograph data were processed offline using PSPHA (De Clerck et al., 2006), which resulted in breath-by-breath information of respiratory timing, respiratory volume, and fraction of end-tidal CO_2 ($FetCO_2$). Results of these analyses were further averaged for each acquisition trial.

Data analysis was carried out using SPSS 22 (IBM, Armonk, NY, United States). Symptom response to CO_2 was defined as a significant within-person difference between symptom ratings

TABLE 2 | Differences in Symptom Report and Respiratory Parameters between CO_2 Responders and Non-responders.

Respiratory	CO_2 Responders		CO_2 Non-responders			
Parameter	**Mean**	***SD***	**Mean**	***SD***	***t****	***P***
Symptom intensity CO_2 (VAS)	14.765	13.506	6.167	8.657	2.67	0.011
Symptom intenity RA (VAS)	2.958	4.810	3.731	5.614	−0.51	0.615
Symptom unpleasantness CO_2 (VAS)	14.530	15.830	5.923	8.488	2.40	0.021
Symptom unpleasantness RA (VAS)	2.529	4.517	3.310	5.590	−0.526	0.602
Expiratory volume CO_2 (1)	0.895	0.326	0.540	0.360	3.51	0.001
Expiratory volume RA (1)	0.789	0.320	0.545	0.392	2.30	0.026
Inspiratory volume CO_2 (1)	0.859	0.340	0.506	0.298	3.79	< 0.001
Inspiratory volume RA (1)	0.752	0.328	0.489	0.275	2.99	0.005
Inspiratory flow CO_2 (1/s)	0.415	0.149	0.282	0.150	3.03	0.004
Inspiratory flow RA (1/s)	0.368	0.131	0.272	0.146	2.34	0.024
Fraction of end-tidal CO_2 (%)	11.22	1.30	9.22	1.83	4.22	< 0.001
Fraction of end-tidal RA (%)	8.95	1.02	7.90	1.34	2.93	0.005
Minute ventilation CO_2 (1/min)	10.739	3.839	7.270	4.048	2.99	0.005
Minute ventilation RA (1/min)	9.407	3.368	6.981	4.068	2.19	0.035

*SD, standard deviation; RA, room air. * df = 46 for VAS ratings, df = 45 for respiratory parameters.*

after the CO_2 trials, compared to room air trials, calculated using independent samples *t*-tests. If participants showed a $p < 0.05$ significant difference in symptom levels either on the symptom intensity or symptom unpleasantness ratings, they were deemed CO_2 responders ($n = 22$). Participants not showing a significant difference were deemed non-responders ($n = 26$). Responders and non-responders did not differ in gender [$X^2(1) = 0.01$, $p = 0.938$], age [$t(46) = -0.17$, $p = 0.864$], negative affectivity [$t(46) = -0.71$, $p = 0.479$], positive affectivity [$t(46) = 0.09$, $p = 0.928$] and fear of suffocation [$t(46) = -0.96$, $p = 0.340$], nor did they differ in assignment to information manipulation groups [$X^2(1) = 0.00$, $p = 1.000$], or assignment of similar/different CS categories during acquisition [$X^2(1) = 0.34$, $p = 0.562$]. Additionally, we explored difference in respiratory parameters of CO_2 responders vs. non-responders. One participant was excluded from these analyses because of equipment failure. In a series of one-way repeated measures ANOVA's, we found significant differences in the two groups for expiratory and inspiratory volume, minute ventilation, inspiratory drive, and $FetCO_2$ and room air and the minute ventilation (**Table 2**).

Using only the responder data, acquisition, retention, and generalization of trigger beliefs was evaluated using multilevel (linear mixed models) analysis. Multilevel models were chosen because these models are less restrictive in variance-covariance assumptions for repeated measures data compared to repeated measures ANOVA, are robust to unbalanced designs, and are less restrictive in the need of having fully nested or fully crossed designs (e.g., clear separation between- and within-subject effects) compared to (repeated measures) ANOVA (Cnaan et al., 1997). Therefore, these models provide an option to deal with the peculiarities of our trigger recognition/generalization dataset (e.g., all participants having CS+ and CS− trials, but having either 20 Gu trials or 10 G+ and 10 G− trials). Models were fitted using random intercepts to account for the data being nested within participants, and were estimated using Maximum Likelihood estimation. SPSS uses Satterthwaite approximation to determine df for *F*-tests/*t*-tests. Model fit was evaluated using Akaike's Information Criterion (AIC). We carried out additional analyses on the full set of participants, while including CO2 responder status as an additional factor. For the CO2 responders, this did not result in major changes to our findings for retention and generalization of trigger beliefs. Results of these analyses are reported as Supplementary Material.

RESULTS

Acquisition of Trigger Beliefs

For acquisition of trigger beliefs, we constructed a multilevel model that included fixed effects of CS (CS+ vs. CS−), Trial (T1–T10), and Trigger Information (general vulnerability vs. specific sensitivities), and included all interactions between these variables. The model also included a random (individual level) effect of CS, with an unstructured variance-covariance matrix. We observed no main effects of CS type [$F(1,22) = 0.260$, $p = 0.615$] nor a CS type × trial interaction [$F(9,396) = 1.190$, $p = 0.300$]. However, this analysis resulted in a significant main effect of trial [$F(9,396) = 5.064$, $p < 0.001$], showing reducing symptom expectancies from the first trial to subsequent trials. This effect was further qualified by Trigger Information [$F(9,396) = 3.066$, $p = 0.001$], showing that this decline in symptom expectancies was specific for participants who had been informed of triggers indicating general vulnerability. The CS × Information interaction was not significant [$F(1,22) = 0.88$, $p = 0.358$], and although the CS × Trial × Information interaction did not reach significance [$F(9,396) = 1.692$, $p = 0.089$], visual inspection of this interaction suggested better differentiation for CS+/CS− symptom expectancies when participants had been given information about triggers as specific sensitivities vs. general vulnerability, (cf. **Figure 1**). Addition of CS category relationship to these analyses did not result in

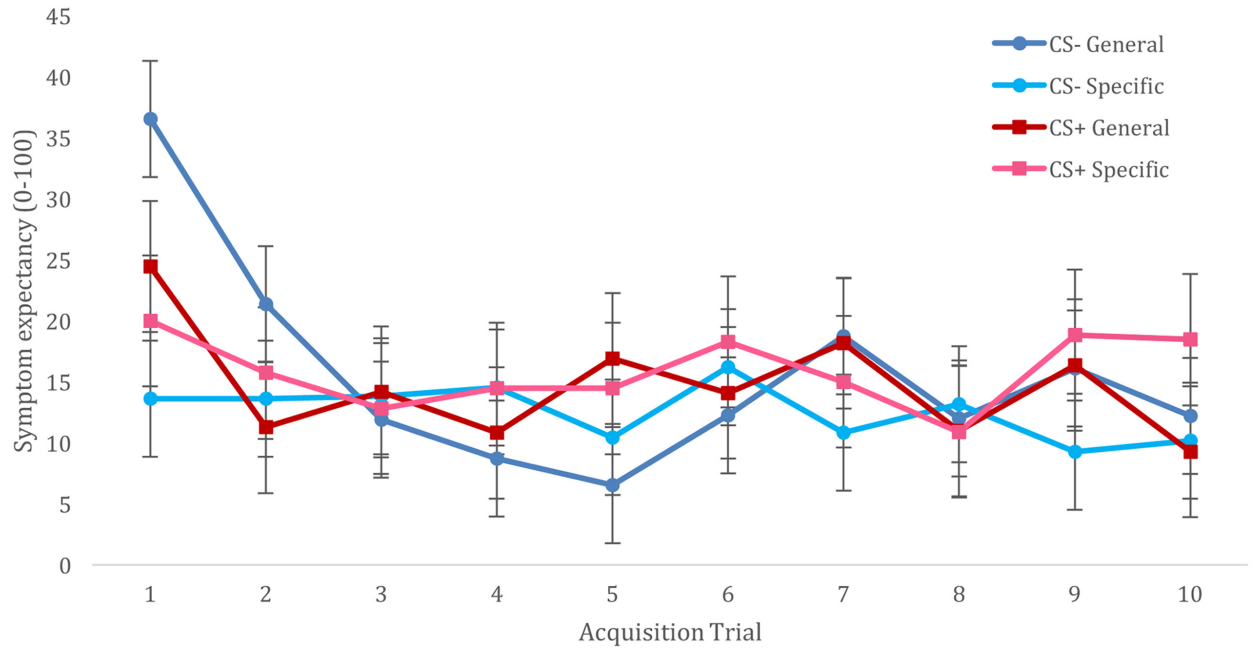


FIGURE 1 | Trigger-symptom expectancies during acquisition phase.

improved model fit (AIC increased from 4504 to 3552) or changes in observed significant effects.

Retention of Trigger Beliefs and Generalization to Novel Exemplars

For retention of trigger beliefs, we constructed a multilevel model that included fixed effects of CS (CS+ vs. CS−), CS novelty (old vs. new), Category Relationship (similar vs. different), and Trigger Information (general vulnerability vs. specific sensitivities), and included all interactions between these variables. The model also included a random (individual level) intercept, to account for the data being nested within participants. Results of this analysis is represented in **Figure 2**. In general, Symptom expectancy was greater for CS+ compared to CS− exemplars [$F(1,858) = 29.094$, $p < 0.001$]. Furthermore, symptom expectancy was greater for old compared to novel trigger exemplars [main effect of CS novelty: $F(1,858) = 11.231$, $p = 0.001$]. This effect was unmodulated by interactions with any of the other model factors, and we observed differential symptom expectancies both for old [$t(858) = 3.812$, $p < 0.001$] as well for novel [$t(858) = 3.995$, $p < 0.001$] category exemplars. Finally, we observed a significant CS × Category Relationship × Trigger Information 3-way interaction [$F(1,858) = 4.174$, $p = 0.041$]. Further exploration of this interaction showed significant differential CS+/CS− expectancies when information was given about general vulnerability and CS categories were more similar [$t(858) = 4.075$, $p < 0.001$] or when information about specific sensitivities was given and CS categories were more different [$t(858) = 5.409$, $p < 0.001$], differences between CS+/CS− for other combinations of Trigger Information and CS Category Relationship were non-significant (but in the expected direction, cf. **Figure 2**). We did not observe any other significant main effects or interactions in this analysis.

Generalization to Novel Trigger Categories

Based on the trigger categories that were used as CS+ and CS−, the novel trigger categories could be coded as G+ (related to CS+), G− (related to CS−) or Gu (unrelated to both CS categories). We constructed a multilevel model that included fixed effects of Stimulus Category (CS+, CS−, G+, G−, Gu), and Trigger Information (general vulnerability vs. specific sensitivities), and included all interactions between these variables. The model also included a random (individual level) intercept, to account for the data being nested within participants. Because of overlap of CS similarity with G categories (cf. **Table 1**), CS similarity was not added as a predictor to this model.

Results showed a main effect of Stimulus Category [$F(4,1745) = 34.832$, $p < 0.001$], further exploration of this effect showed that CS+ ratings were significantly higher compared to all other categories [CS+/CS− $t(1738) = 7.824$, $p < 0.001$; CS+/Gu $t(1751) = 9.588$, $p < 0.001$; CS+/G− $t(1746) = 7.044$, $p < 0.001$; CS+/G+ $t(1746) = 7.011$, $p < 0.001$], and that symptom expectancies for CS− exemplars were higher compared to Gu symptom expectancies [$t(1751) = 3.329$, $p = 0.009$], but not different from G+ and G− symptoms expectancies [CS−/G+ $t(1746) = 1.165$, $p > 0.99$; CS−/G− $t(1746) = 1.198$, $p > 0.99$]. Gu, G+, and G− ratings did not differ from each other (cf. **Figure 3**).

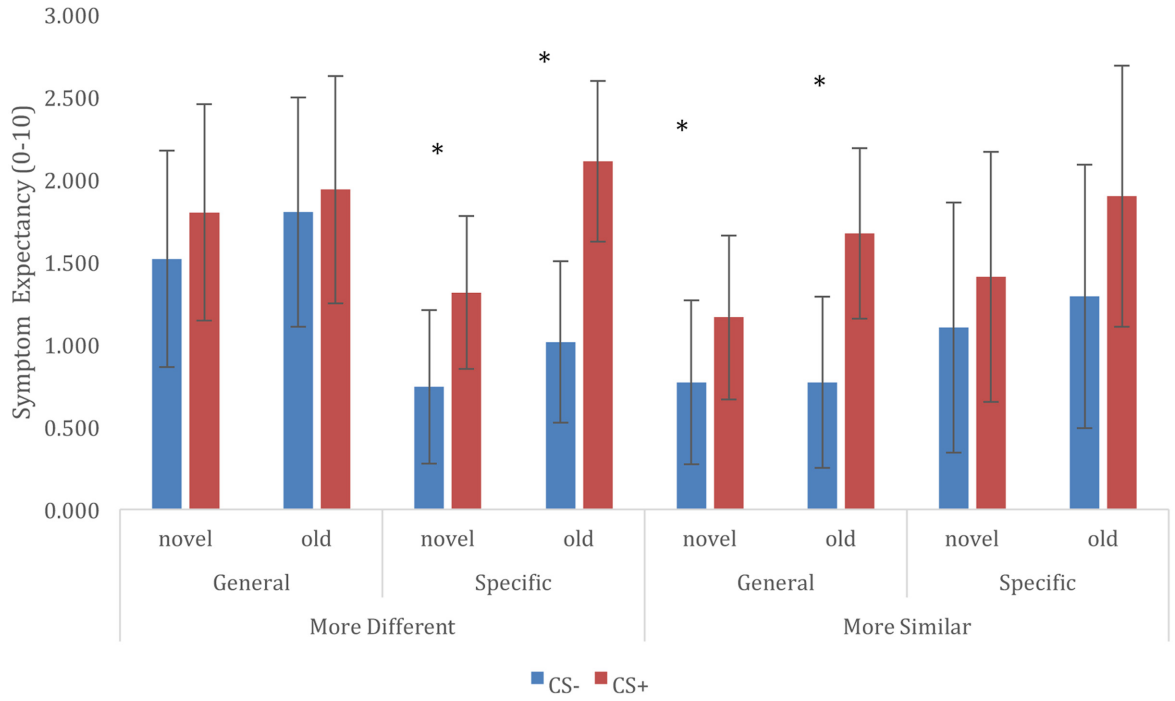


FIGURE 2 | Retention of trigger-symptom expectancies, and generalization to novel trigger exemplars, depending on conditioned stimuli (CS), Trigger Information, and CS Category Relationship. * CS+/CS– symptom expectancy differ at $p < 0.05$.

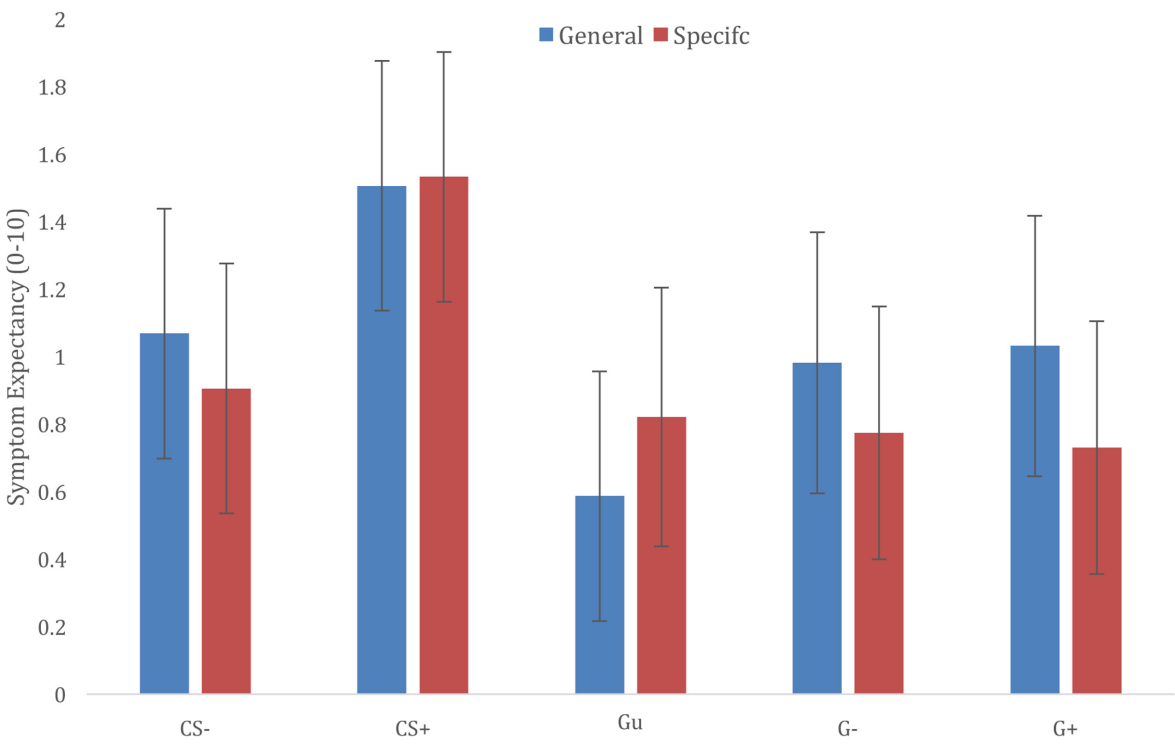


FIGURE 3 | Generalization of trigger-symptom contingencies according to different information groups (general vulnerability vs. specific sensitivities).

The main effect of Trigger Information was not significant [$F(1,22) = 0.025$, $p = 0.875$], nor did the Trigger Information × Stimulus Category interaction yield a significant effect [$F(4,1745) = 1.951$, $p = 0.100$]. Visual exploration of the interaction suggested that providing information about specific sensitivities may prevent generalization to generalization categories that were related to CS categories (G+; G−), but not Gu category triggers (cf. **Figure 3**).

DISCUSSION

In this experiment, we used a laboratory analog task in order to investigate the impact of information about the causal structure of asthma triggers and symptoms (asthma triggers being an indication of general vulnerability vs. specific sensitivities) on the acquisition, retention, and generalization of category-based trigger-symptom contingencies.

Results of the acquisition phase did not show clear evidence for the acquisition of category based symptom expectancies. This lack of clear acquisition effects is contrary to previous results with a similar experimental method, in a study that did not include explicit information about general vulnerability or specific sensitivities (Janssens et al., 2015). This may suggest that both types of information hinder the acquisition of differential trigger expectancies, although the large number of non-responders in the current experiment may limit the value of this comparison (cf. supra).

During the retention and generalization phase, we did observe retention of category based symptom expectancies, and generalization of these expectancies to novel CS+/CS− category exemplars. Information about specific sensitivities or general vulnerability had an impact on symptom expectancies, but this effect was not straightforward, as it was moderated by the trigger category relationship. Information about specific sensitivities led to better retention of differential expectancies when CS Categories had been more different, whereas information about general vulnerability led to better retention when CS Categories had been more similar. At first sight, the emergence of differential symptom expectancies after the acquisition phase may be puzzling. However, it is possible that the abstraction of category level information from the unique exemplars does not happen right away, and therefore would not show up on the trial by trial expectancy ratings. Furthermore, previous studies have shown category level consolidation effects, extending to other CS+ exemplars (Dunsmoor et al., 2015), which could explain why we do find differential CS+/CS− retention effects in absence clear differential learning during acquisition.

When confronted with novel (generalization) trigger categories, we could not confirm our hypothesis that trigger expectancies generalize to trigger categories that are related to the CS categories. However, participants did show some selectivity in generalization to novel categories, as evidenced by our finding that symptom expectancies for Gu triggers were lower than expectancies for CS+ and CS− triggers. Interestingly, we did not observe any differences between G+ and G− category exemplars, although the limited number of participants precludes us from making strong inferences about this. Generalization to novel categories was not moderated by our information manipulation, although visual inspection of the results was in line with information about asthma being caused by a general vulnerability leading to stronger symptom expectancies for trigger categories that were similar to CS+ or CS− categories, but not to potential triggers from unrelated categories.

Despite the many interaction effects that we observed, the results in the different conditions of our experiment demonstrate the impact of prior information on the acquisition, retention, and generalization of category-based trigger-symptom contingency beliefs. As the information conditions in our experiment mimic aspects of trigger-related information or advice that is given to patients by physicians or in internet-based asthma information, our findings may be of relevance to the management of asthma in daily life, as they suggest that experience-based beliefs about asthma triggers are shaped by prior information about asthma causality, as well as individual differences in symptom perception. The effects of prior information on generalization of trigger beliefs may be especially relevant, as they may help to explain the individual differences in asthma trigger beliefs that have been observed in individuals with asthma (Ritz et al., 2016) and associated differences in trigger avoidance strategies (Vernon et al., 2012).

Limitations

Our findings are limited by the observation that less than half of participants responded in a consistent way to our symptom induction of 60 s inhalation of an air mixture containing 7.5% CO_2. Although previous studies had used longer inhalation periods (ranging from 90 s to 20 min) of 7.5% CO_2 air mixtures in order to induce respiratory symptoms or symptoms of anxiety (Bailey et al., 2005; Bogaerts et al., 2005; Pappens et al., 2012; Janssens et al., 2015), our decision to use shorter duration symptom trials was motivated by a perceived need to reduce symptom burden (participation time), and did occur after pilot testing suggesting that participants were able to differentiate between 60 s room air and CO_2 inhalation. Nevertheless, the results of this study show that longer periods of CO_2 inhalation may be needed to reduce variability in symptom response and increase the differences between inhalation of a 7.5% CO_2 air mixture and room air inhalation. Furthermore, even if participants reliably responded differently to the 7.5% CO_2 air mixture and room air, they may not have picked up on these differences in a way that would lead them to form clear symptom-trigger contingencies. In our previous experiment using 90 s inhalation of the 7.5% CO_2 air mixture, differentiation between CS+ and CS− symptom expectancies was markedly better.

Furthermore, our findings are limited in that we did not test behavioral outcomes related to these generalized trigger beliefs, nor did we test if these generalized triggers were sensitive to disconfirmation. However, studies on fear generalization have shown that generalization of negative outcome expectancies is accompanied by increased physiological manifestations of fear, as well as increased avoidance behavior to the generalization stimuli (van Meurs et al., 2014; Dymond et al., 2015), suggesting that generalized trigger-symptom contingencies can

have an impact on trigger related behaviors. Nevertheless, future studies investigating effects of extinction on generalized trigger-symptom beliefs are needed to further gauge the impact that generalization can have in this domain.

A final limitation – as in many lab-based studies – is that design decisions that were aimed at improving internal validity may have reduced external validity of our experimental design. As noted in our previous study (Janssens et al., 2015), the use of uncommon allergens as experimental asthma triggers helps to isolate specific aspects of triggers as a potential basis of trigger acquisition and generalization, but these aspects may differ from the types of potential triggers that are experienced in real life. Similarly, the selection participants that do not have a history of allergy helps us to mimic conditions that parallel an early phase of asthma trigger identification, but may preclude generalization the lived and contextualized experience of individuals with asthma that may use a variety of information to infer trigger-symptom contingencies (Caress et al., 2002; Vernon et al., 2013).

CONCLUSION

Our findings show that information about causality in asthma and knowledge about conceptual relationships between trigger categories influences the retention of category-based differential trigger-symptom expectancies, and generalization of these expectancies to novel trigger exemplars. Furthermore, retention and generalization of symptom expectancies was moderated by the similarity of CS+/CS− as well as similarities between CS and G categories. These findings underscore the role of higher order cognitions in contingency learning, and may help us to understand individual differences in asthma trigger beliefs that emerge over time. Finally, our findings suggest that pre-existing beliefs about asthma and asthma triggers may need to be taken into account when informing individuals with asthma about asthma trigger identification as an asthma management strategy, as these beliefs may impact subsequent learning of trigger-symptom contingencies in individuals with asthma.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of ethical guidelines of the American Psychological Association (APA) with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Medical Ethical Committee of 'University Hospitals Leuven.'

AUTHOR CONTRIBUTIONS

TJ, IVD, and OVdB conceived the research questions and methodology, EC and TJ conducted the research and analyzed the data, EC, TJ, and OVdb contributed to writing the manuscript.

FUNDING

This research was funded by a Postdoctoral Fellowship awarded to TJ by the Flemish Research Foundation (Fonds Wetenschappelijk Onderzoek – Vlaanderen).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00926/full#supplementary-material

REFERENCES

Ahmed, O., and Lovibond, P. F. (2015a). The impact of instructions on generalization of conditioned fear in humans. *Behav. Ther.* 46, 597–603. doi: 10.1016/j.beth.2014.12.007

Ahmed, O., and Lovibond, P. F. (2015b). The impact of previously learned feature-relevance on generalisation of conditioned fear in humans. *J. Behav. Ther. Exp. Psychiatry* 46, 59–65. doi: 10.1016/j.jbtep.2014.08.001

Bailey, J. E., Argyropoulos, S. V., Kendrick, A. H., and Nutt, D. J. (2005). Behavioral and cardiovascular effects of 7.5% CO2 in human volunteers. *Depress. Anxiety* 21, 18–25. doi: 10.1002/da.20048

Bogaerts, K., Notebaert, K., Van Diest, I., Devriese, S., De Peuter, S., and Van den Bergh, O. (2005). Accuracy of respiratory symptom perception in different affective contexts. *J. Psychosom. Res.* 58, 537–543. doi: 10.1016/j.jpsychores.2004.12.005

Bousquet, J., Schünemann, H. J., Samolinski, B., Demoly, P., Baena-Cagnani, C. E., Bachert, C., et al. (2012). Allergic rhinitis and its impact on asthma (ARIA): achievements in 10 years and future needs. *J. Allergy Clin. Immunol.* 130, 1049–1062. doi: 10.1016/j.jaci.2012.07.053

Brough, H. A., Turner, P. J., Wright, T., Fox, A. T., Taylor, S. L., Warner, J. O., et al. (2015). Dietary management of peanut and tree nut allergy: what exactly should patients avoid? *Clin. Exp. Allergy* 45, 859–871. doi: 10.1111/cea.12466

Caress, A.-L., Luker, K., Woodcock, A., and Beaver, K. (2002). An exploratory study of priority information needs in adult asthma patients. *Patient Educ. Couns.* 47, 319–327. doi: 10.1016/S0738-3991(02)00005-8

Cnaan, A., Laird, N., and Slasor, P. (1997). Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat. Med.* 16, 2349–2380. doi: 10.1002/(SICI)1097-0258(19971030)16:20<2349::AID-SIM667>3.0.CO;2-E

Croft, D. R., and Peterson, M. W. (2002). AN evaluation of the quality and contents of asthma education on the world wide web*. *Chest* 121, 1301–1307. doi: 10.1378/chest.121.4.1301

De Clerck, A., Verschuere, B., Crombez, G., and De Vlieger, P. (2006). Psychophysiological analysis (PSPHA): a modular script based program for analyzing psychophysiological data. *Behav. Res. Methods* 38, 504–510. doi: 10.3758/BF03192805

De Peuter, S., Lemaigre, V., Van Diest, I., and Van den Bergh, O. (2008). Illness-specific catastrophic thinking and overperception in asthma. *Health Psychol.* 27, 93–99. doi: 10.1037/0278-6133.27.1.93

Dunsmoor, J. E., Martin, A., and LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biol. Psychol.* 89, 300–305. doi: 10.1016/j.biopsycho.2011.11.002

Dunsmoor, J. E., Mitroff, S. R., and LaBar, K. S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learn. Mem.* 16, 460–469. doi: 10.1101/lm.1431609

Dunsmoor, J. E., and Murphy, G. L. (2015). Categories, concepts, and conditioning: how humans generalize fear. *Trends Cogn. Sci.* 19, 73–77. doi: 10.1016/j.tics.2014.12.003

Dunsmoor, J. E., Murty, V. P., Davachi, L., and Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature* 520, 345–348. doi: 10.1038/nature14106

Dymond, S., Dunsmoor, J. E., Vervliet, B., Roche, B., and Hermans, D. (2015). Fear generalization in humans: systematic review and implications for anxiety disorder research. *Behav. Ther.* 46, 561–582. doi: 10.1016/j.beth.2014.10.001

Engelen, U., De Peuter, S., Victoir, A., Van Diest, I., and Van den Bergh, O. (2006). Verdere validering van de "Positive and Negative Affect Schedule" (PANAS) en vergelijking van twee Nederlandstalige versies (Further validation of the positive and negative affect schedule (PANAS) and comparison of two Dutch versions). *Gedrag Gezondheid* 34, 89–102.

Global Initiative for Asthma (GINA) (2016). *Global Strategy for Asthma Management and Prevention*. Available at: http://ginasthma.org/wp-content/uploads/2016/04/wms-GINA-2016-main-report-final.pdf

Hagger, M. S., and Orbell, S. (2003). A meta-analytic review of the common-sense model of illness representations. *Psychol. Health* 18, 141–184. doi: 10.1080/0887044031000813 21

Halm, E. A., Mora, P., and Leventhal, H. (2006). No symptoms, no asthma: the acute episodic disease belief is associated with poor self-management among inner-city adults with persistent asthma. *Chest* 129, 573–580. doi: 10.1378/chest.129.3.573

Horne, R., and Weinman, J. (1999). Patients' beliefs about prescribed medicines and their role in adherence to treatment in chronic physical illness. *J. Psychosom. Res.* 47, 555–567. doi: 10.1016/S0022-3999(99)00057-4

Huckvale, K., Car, M., Morrison, C., and Car, J. (2012). Apps for asthma self-management: a systematic assessment of content and tools. *BMC Med.* 10:144. doi: 10.1186/1741-7015-10-144

Janssens, T., and Harver, A. (2015). Effects of symptom perception interventions on trigger identification and quality of life in children with asthma. *Pulm. Med.* 2015:825137. doi: 10.1155/2015/825137

Janssens, T., Martens, F., Storms, N., and Van den Bergh, O. (2015). Generalization of respiratory symptom triggers. *Behav. Ther.* 46, 689–698. doi: 10.1016/j.beth.2015.05.003

Janssens, T., and Ritz, T. (2013). Perceived triggers of asthma: key to symptom perception and management. *Clin. Exp. Allergy* 43, 1000–1008. doi: 10.1111/cea.12138

Janssens, T., Verleden, G., De Peuter, S., Petersen, S., and Van den Bergh, O. (2011). The influence of fear of symptoms and perceived control on asthma symptom perception. *J. Psychosom. Res.* 71, 154–159. doi: 10.1016/j.jpsychores.2011.04.005

Janssens, T., Verleden, G., De Peuter, S., Van Diest, I., and Van den Bergh, O. (2009). Inaccurate perception of asthma symptoms: a cognitive-affective framework and implications for asthma treatment. *Clin. Psychol. Rev.* 8, 211–219. doi: 10.1016/j.cpr.2009.02.006

Kaptein, A. A., Klok, T., Moss-Morris, R., and Brand, P. L. (2010). Illness perceptions: impact on self-management and control in asthma. *Curr. Opin. Allergy Clin. Immunol.* 10, 194–199. doi: 10.1097/ACI.0b013e32833950c1

Leventhal, H., Meyer, D., and Nerenz, D. (1980). The common sense representation of illness danger. *Contribut. Med. Psychol.* 2, 7–30.

Li, J. T. C., Andrist, D., Bamlet, W. R., and Wolter, T. D. (2000). Accuracy of patient prediction of allergy skin test results. *Ann. Allergy Asthma Immunol.* 85, 382–384. doi: 10.1016/S1081-1206(10)62550-1

McQuaid, E. L., Howard, K., Kopel, S. J., Rosenblum, K., and Bibace, R. (2002). Developmental concepts of asthma: reasoning about illness and strategies for prevention. *J. Appl. Dev. Psychol.* 23, 179–194. doi: 10.1016/S0193-3973(02)00103-X

Pappens, M., De Peuter, S., Vansteenwegen, D., Van den Bergh, O., and Van Diest, I. (2012). Psychophysiological responses to CO2 inhalation. *Int. J. Psychophysiol.* 84, 45–50. doi: 10.1016/j.ijpsycho.2012.01.008

Peters, S. P., Jones, C. A., Haselkorn, T., Mink, D. R., Valacer, D. J., and Weiss, S. T. (2007). Real-world evaluation of asthma control and treatment (REACT): findings from a national web-based survey. *J. Allergy Clin. Immunol.* 119, 1454–1461. doi: 10.1016/j.jaci.2007.03.022

Rabe, K. F., Adachi, M., Lai, C. K. W., Soriano, J. B., Vermeire, P. A., Weiss, K. B., et al. (2004). Worldwide severity and control of asthma in children and adults: the global asthma insights and reality surveys. *J. Allergy Clin. Immunol.* 114, 40–47. doi: 10.1016/j.jaci.2004.04.042

Ritz, T., Steptoe, A., Bobb, C., Harris, A. H. S., and Edwards, M. (2006). The asthma trigger inventory: validation of a questionnaire for perceived triggers of asthma. *Psychosom. Med.* 68, 956–965. doi: 10.1097/01.psy.0000248898.59557.74

Ritz, T., Wittchen, H.-U., Klotsche, J., Mühlig, S., and Riedel, O. (2016). Asthma trigger reports are associated with low quality of life, exacerbations, and emergency treatments. *Ann. Am. Thorac. Soc.* 13, 204–211. doi: 10.1513/AnnalsATS.201506-390OC

Smith, H., Gooding, S., Brown, R., and Frew, A. (1998). Evaluation of readability and accuracy of information leaflets in general practice for patients with asthma. *BMJ* 317, 264–265. doi: 10.1136/bmj.317.7153.264

Smith, H. E., Hogger, C., Lallemant, C., Crook, D., and Frew, A. J. (2009). Is structured allergy history sufficient when assessing patients with asthma and rhinitis in general practice? *J. Allergy Clin. Immunol.* 123, 646–650. doi: 10.1016/j.jaci.2008.11.005

Spruyt, A., Clarysse, J., Vansteenwegen, D., Baeyens, F., and Hermans, D. (2010). Affect 4.0: a free software package for implementing psychological and psychophysiological experiments. *Exp. Psychol.* 57, 36–45. doi: 10.1027/1618-3169/a000005

Trollvik, A., and Severinsson, E. (2004). Parents' experiences of asthma: process from chaos to coping. *Nursing Health Sci.* 6, 93–99. doi: 10.1111/j.1442-2018.2004.00179.x

Van Diest, I., Smits, D., Decremer, D., Maes, L., and Claes, L. (2010). The Dutch claustrophobia questionnaire: psychometric properties and predictive validity. *J. Anxiety Disord.* 24, 715–722. doi: 10.1016/j.janxdis.2010.05.003

van Meurs, B., Wiggert, N., Wicker, I., and Lissek, S. (2014). Maladaptive behavioral consequences of conditioned fear-generalization: a pronounced, yet sparsely studied, feature of anxiety pathology. *Behav. Res. Ther.* 57, 29–37. doi: 10.1016/j.brat.2014.03.009

Vernon, M. K., Bell, J. A., Wiklund, I., Dale, P., and Chapman, K. (2013). Asthma control and asthma triggers: the patient perspective. *J. Asthma Allergy Educ.* 4, 155–164. doi: 10.1177/2150129713483307

Vernon, M. K., Wiklund, I., Bell, J. A., Dale, P., and Chapman, K. R. (2012). What do we know about asthma triggers? A review of the literature. *J. Asthma* 49, 991–998. doi: 10.3109/02770903.2012.738268

Vervliet, B., and Geens, M. (2014). Fear generalization in humans: impact of feature learning on conditioning and extinction. *Neurobiol. Learn. Mem.* 113, 143–148. doi: 10.1016/j.nlm.2013.10.002

Vervliet, B., Kindt, M., Vansteenwegen, D., and Hermans, D. (2010). Fear generalization in humans: impact of verbal instructions. *Behav. Res. Ther.* 48, 38–43. doi: 10.1016/j.brat.2009.09.005

von Leupoldt, A., and Dahme, B. (2012). Looking at allergens increases symptom report in patients with allergic asthma. *J. Asthma* 49, 1027–1029. doi: 10.3109/02770903.2012.733994

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Conflict of Interest Statement: 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.

Copyright © 2017 Janssens, Caris, Van Diest and Van den Bergh. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 February 2017
doi: 10.3389/fpsyg.2017.00077

How You Move Is What I See: Planning an Action Biases a Partner's Visual Search

Dominik Dötsch[1], Cordula Vesper[2] and Anna Schubö[1]*

[1] Cognitive Neuroscience of Perception and Action, Faculty of Psychology, Philipp University of Marburg, Marburg, Germany,
[2] Department of Cognitive Science, Central European University, Budapest, Hungary

Activating action representations can modulate perceptual processing of action-relevant dimensions, indicative of a common-coding of perception and action. When two or more agents work together in joint action, individual agents often need to consider not only their own actions and their effects on the world, but also predict the actions of a co-acting partner. If in these situations the action of a partner is represented in a functionally equivalent way to the agent's own actions, one may also expect interaction effects between action and perception across jointly acting individuals. The present study investigated whether the action of a co-acting partner may modulate an agent's perception. The "performer" prepared a grasping or pointing movement toward a physical target while the "searcher" performed a visual search task. The performer's planned action impaired the searcher's perceptual performance when the search target dimension was relevant to the performer's movement execution. These results demonstrate an action-induced modulation of perceptual processes across participants and indicate that agents represent their partner's action by employing the same perceptual system they use to represent an own action. We suggest that task representations in joint action operate along multiple levels of a cross-brain predictive coding system, which provides agents with information about a partner's actions when they coordinate to reach a common goal.

Keywords: joint action, task representations, action-perception links, visual attention, intentional weighting, predictive coding

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Roland Thomaschke, University of Regensburg, Germany
Michael Ziessler, Liverpool Hope University, UK

****Correspondence:***
Dominik Dötsch
dominik.doetsch@uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *29 September 2016*
Accepted: *12 January 2017*
Published: *07 February 2017*

Citation:
Dötsch D, Vesper C and Schubö A (2017) How You Move Is What I See: Planning an Action Biases a Partner's Visual Search. Front. Psychol. 8:77. doi: 10.3389/fpsyg.2017.00077

INTRODUCTION

Few activities in our everyday life are performed in isolation, without considering another person's actions. Instead, when people act together to reach a common goal in joint action, individual agents tend to represent not only their own part of the task, but often also form a cognitive representation of their partner's part. Agents may use these representations to successfully coordinate with their partner (Vesper et al., 2010). However, the influence of a co-acting partner on an agent's performance is not limited to situations in which the partner's response needs to be considered to fulfill the own part of the task. In fact, evidence for a modulation of task performance in joint action was initially found in response time (RT) paradigms in which representing the partner's task could be detrimental to own task performance. In these studies, two complementary tasks originally performed by one participant were split between two co-acting participants. For instance, in the joint Simon task (Sebanz et al., 2003), two participants sitting side-by-side performed a Go–Nogo

task that also included a task-irrelevant spatial stimulus. Compatibility between the spatial stimulus and the responding agent's location affected RTs. RTs were slower when the spatial stimulus pointed toward the partner, similar to the results found in individual agents when stimulus and response location did not match (Simon and Rudell, 1967). This joint Simon effect has been interpreted as the consequence of an activation of the representation of the partner's task, leading to interference during selection of the agent's own response.

Observing another person performing an action was also found to influence own performance. According to ideomotor theories, observing another person's action activates the same representations in the observer's cognitive system that is usually employed to produce an own action (Prinz, 1990; Hommel et al., 2001; Hommel, 2009). Behavioral studies support this view, as they have shown that observing movements compatible to the own task facilitates, while observing incompatible movements impedes task performance (Brass et al., 2001). The physiological basis of these compatibility effects was described as a motor resonance (Sebanz et al., 2006; Ménoret et al., 2013) implying that response-relevant motor regions are pre-activated by observing compatible movements and response-irrelevant motor regions have to be suppressed when observing incompatible movements. Indeed, similar activations have been recorded both in human and primate motor areas of the brain during action perception and during action execution (di Pellegrino et al., 1992; Rizzolatti and Craighero, 2004; Fadiga et al., 2005; Newman-Norlund et al., 2008; Bekkering et al., 2009).

Action simulation plays an important role in predicting another agent's movements, for example, when an observed action is temporarily occluded (Springer et al., 2013). Action simulation is thus not only based on perception and subsequent mapping of movement, but on the creation of goal-directed action predictions. The precision of such predictions depends on the level of motor experience with the movement (Cross et al., 2006; Güldenpenning et al., 2013). Representations of a partner's movement can include the movement's biomechanical and sensory consequences, as agents were found to adapt their own movements to increase their partner's postural comfort at the movement goal (Dötsch and Schubö, 2015), similar to what is known from individual agents maximizing their own end-state comfort (Cohen and Rosenbaum, 2004).

The above studies demonstrate the influence of a partner's task on different levels of own task processing including response selection, motor planning, and movement execution. Another process susceptible to the influence of a partner's task in joint action is visual attention. For instance, Baess and Prinz (2015) used a joint Go–Nogo task in which a first cue identified which agent had to respond, while a second cue signaled which response was required. Thereby, agent identification was disentangled from response selection. Results showed that the N1, an ERP component indicative of early perceptual processing, was less pronounced in the joint compared to the single action condition for physically identical agent identification cues. This implies that the early stage of perceptual processing was modulated by the representation of the partner's task. Joint action thus not only influences agents on the level of response selection as in the joint Simon task, but can change the way agents perceive their environment.

In a series of experiments, Wykowska et al. (2009, 2012; Wykowska and Schubö, 2012) demonstrated that action planning can directly affect perceptual processing of action-relevant dimensions. In their paradigm, individual participants had to prepare a movement that had to be executed later in the trial at the onset of a Go signal. During movement preparation, participants performed a visual search task. Only after completion of the search task, a cue indicated the goal of the prepared movement. Results in the search task showed that RTs differed depending on the congruency between the prepared movement and the dimension in which the search target differed from the distractors: Preparing a grasping movement facilitated the detection of size targets, resulting in faster RTs compared to trials in which a pointing movement had to be prepared. Preparing a pointing movement accelerated RTs to luminance targets compared to when a grasping movement had to be prepared. This modulation of perceptual processing by a planned action has been interpreted in terms of intentional weighting (Hommel et al., 2001; Hommel, 2009; Memelink and Hommel, 2013). Similar to the ideomotor theory, this account relies on the idea that actions are represented by their sensory consequences in a common-coding format of perception and action (Prinz, 1997; Hommel et al., 2001; Hommel, 2009). According to intentional weighting, action planning results in prioritized processing of those perceptual dimensions that are delivering information relevant to achieve the intended action goal. To optimally adjust open action parameters, the perceptual system preferably processes those dimensions that are relevant to specify and execute the action (Wykowska et al., 2012). For example, grasping an object requires adjusting the grip aperture to the size of the object, while other perceptual dimensions such as the object's color are irrelevant. Thus when a grasping movement had to be prepared, perceptual processing of the size dimension was prioritized in the intermediate search task, resulting in faster target detection than when a pointing movement was prepared. Several other studies have shown facilitation of the perception of action-relevant dimensions. Planning a grasping movement was reported to facilitate the detection of orientation targets compared to a pointing movement (Bekkering and Neggers, 2002). Similarly, preparing a precision grip facilitated the perception of a change in small objects in a change blindness test, while a power grip facilitated the perception of a change in larger objects (Symes et al., 2008). Furthermore, grasping movements were initiated faster when a Go cue was oriented similar to the orientation of the goal object compared to a differently oriented cue, indicating faster processing of stimuli sharing perceptual features with the action goal (Craighero et al., 1999). In the paradigm of Wykowska and colleagues, the P1 component, an ERP correlate of early sensory processing, was larger for luminance targets when participants prepared a pointing compared to a grasping movement. For size targets, the N2pc component was larger when preparing a grasping compared to a pointing movement (Wykowska and Schubö, 2012).

In other paradigms, however, action planning impaired the perception of stimuli congruent to the planned action. For example, Müsseler and Hommel (1997a) asked participants to prepare a left or right button press. Before executing the keypress response, a left or right pointing arrow had to be identified. The probability of correctly identifying the arrow was lower when a congruent response was prepared compared to an incongruent response. Similar observations were made using a detection task (Müsseler and Hommel, 1997b). The authors concluded that planning an action leads to a temporary blindness to stimuli that resemble the anticipated sensory consequences of the planned action. They suggested that this blindness prevents that the sensory consequences of the executed action activate the same action plan again in the common-coding system. The temporary blindness thus averts recurring action-perception loops.

To account for both facilitation and impairment of perception by action, Thomaschke (2012; Thomaschke et al., 2012) suggested a planning and control model (PCM) of motorvisual priming. The PCM assumes that there are two distinct systems of action planning and movement control. These systems work together to select actions and control their execution. The action planning system primarily processes categorical action representations, e.g., which response is required and which effector is used for response execution (e.g., a right hand grasping movement). The movement control system adjusts specific parameters of movement execution (e.g., the grip aperture needed for grasping). According to PCM, actions impair or facilitate perception depending on whether the action can fully be specified in advance, or whether it requires online adjustment of open parameters. Impairment of perception is observed when the action planning system "binds" representations of (perceptual) features of the planned action. Feature dimensions bound by movement planning are less available to other processes (e.g., perception). PCM suggests that this binding shields the planned action from other cognitive processes to ensure its successful execution. Facilitation of perception, on the other hand, results when an action requires online adjustment of movement parameters in the movement control system (Glover, 2004). In this case, those perceptual dimensions are preferably processed that deliver information for adjusting open action parameters (see also Wykowska et al., 2012; Memelink and Hommel, 2013).

The objective of the present study was to extend and combine previous work on the interaction of action and perception in single and joint action. Specifically, the aim was to test whether a partner's action planning modulates an agent's perceptual processing in a joint action task similarly to what is known from individual dual task performance. To this end, the paradigm used by Wykowska et al. (2009) was adapted for two co-acting participants sitting side-by-side. In particular, one participant (the "performer") had to prepare a pointing or grasping movement while the other participant (the "searcher") searched for a size or a luminance target in a search display. If the searcher represented the performer's movement task similar to an own movement, relying on the common-coding format of perception and action, we assumed that the searcher would not only represent features of the own visual search task but additionally include features relevant to the performer's movement. Consequently, the searcher's perceptual processing should be modulated depending on the congruency between the dimension relevant to the performer's prepared movement and the search target dimension. Trials were considered congruent when the searcher had to detect a luminance target while the performer prepared a pointing movement, and when the searcher had to detect a size target while the performer prepared a grasping movement. Incongruent trials had reversed search target-movement task assignment.

Based on previous studies, two possible modulations by action-perception congruency can be assumed: On the one hand, the modulation may take the form of facilitated responses in the search task (shorter RTs, higher response accuracy) in congruent compared to incongruent trials as observed in the single agent version of the paradigm (Wykowska et al., 2009, 2012; Wykowska and Schubö, 2012). On the other hand, as described above, previous research indicates an interfering influence of a partner's task (e.g., in the joint Simon task, Sebanz et al., 2003). Representing the performer's task may impose an additional load upon searchers' perceptual system, resulting in impeded search task performance (longer RTs, lower response accuracy) in congruent compared to incongruent trials.

Our main research question focused on the modulation of the searcher's task performance by the performer's movement planning. In addition, we investigated the influence of the searcher's perceptual task on the performer's movement execution. Tracking the motion of the performer's thumb and index finger allowed investigating whether the congruency between the searcher's target dimension and the dimension relevant to the performer's movement also influenced movement performance.

MATERIALS AND METHODS

Participants

Sixty-six volunteers (39 female, 27 male; mean age 22.9 years) were grouped into 33 pairs. One participant was excluded because she had an accuracy of only 35% in size target absent trials in the search task [overall mean accuracy for these targets (*SD*) 84.8 (12.6)%]. All 65 remaining participants (38 female, 27 male; mean age 22.9 years) were right handed (mean laterality quotient 76 in the Edinburgh Handedness Inventory, Oldfield, 1971) and had normal or corrected-to-normal visual acuity (tested with a Binoptometer 3, Oculus, Germany).

Stimuli

Stimuli were presented on a 22-inch NT-TFT display (Syncmaster 2233, Samsung, Korea) with a 100 Hz refresh rate placed centrally between participants sitting side-by-side at a distance of 100 cm from the screen. Stimulus presentation and the experimental procedure were controlled by E-Prime 2.0.8 (Psychology Software Tools, Inc., USA) running on a Windows 7 computer.

Search Task

The search display (**Figures 1A,B**) contained 28 items (gray circles of 1.2° of visual angle; 15 cd/m^2 of luminance, measured 100 cm centrally in front of the screen with an Konica Minolta LS-100 spectrometer) positioned on three concentric imaginary circles with diameters of 5.2°, 9.1°, and 13.4° around the fixation cross on a white background (132 cd/m^2). Item positions on the outer two circles were equidistant around the imaginary circles and mirror-symmetric, the four positions on the inner circle were offset from the cardinal axes by 22.5° and were mirrored along the vertical axis in half of all displays. The target was presented on one of four positions in the upper left/right or lower left/right on the middle circle (indicated by dotted circles in **Figure 1**) in half of the trials. The target either differed in luminance (lighter gray: 58 cd/m^2) or in size (larger circle: 1.6°) from the rest of the items in the search display.

Movement Task

The movement cue (**Figures 1C,D**) consisted of a black and white photo of a hand performing either a grasping or a pointing movement toward a medium sized movement object in the medium shade of gray (see "Apparatus" section). The depicted object was centered on screen while the hand and part of the arm extended toward the lower left of the screen 12° off center. Both the pointing and the grasping cue were of the same average brightness (109 cd/m^2). The Go cue consisted of the text "GO 1," "GO 2," or "GO 3" sized 2.3° by 0.75°. It was presented 1° below the horizontal midline of the screen, either 10° to the left, centrally, or 10° to the right of the vertical midline, depending on the position of the object relative to the performer.

Apparatus

Participants were seated side-by-side in comfortable chairs in a dimly lit, sound attenuated room. The performer was sitting on the left and performed the movement task with the left hand. The searcher was sitting on the right and responded to the search display with the right hand (**Figure 2**). Participants were instructed to keep their inactive hand on their thigh. A starting position for the movement task was marked by a cross on a button plate embedded in the middle of a board positioned over the left chair's armrests, 80 cm in front of the screen. Performers were asked to keep their left thumb and index finger on this position until movement execution, depressing the button plate. Searchers responded to the search display by pressing one of two buttons with their right index and middle finger on a response box fixated on their right thigh near the knee with a belt. In front of the performer, three objects were placed as targets for the movement task. The objects were 8 cm high plastic cylinders mounted on stands facing the display. There was always one big (diameter of 8 cm), one medium sized (diameter of 6 cm), and one small object (diameter of 4 cm) present. One of the objects was always a dark shade of gray, one was a medium shade of gray, and one was a light shade of gray (1.4, 0.6, and 0.2 cd/m^2 of luminance under experimental lighting conditions from the performer's viewing distance, respectively). The left and right objects were positioned 46–53 cm in front of the screen, the middle object was positioned 42–49 cm in front of the screen. At the beginning of the experiment, a comfortable distance (in 1 cm steps) was determined at which participants could reach all objects without moving in their chair. This setting was kept the same for all objects for each participant via markings on the table.

The performer's movements were recorded using a magnetic motion tracking device (Polhemus Liberty 240/8, Polhemus Inc., USA) measuring six degrees of freedom (X, Y, and Z position and three rotational angles) at a sampling rate of 240 Hz. Tracking sensors were attached on top of the performer's left thumb and index finger with plaster tape, aligned with the end of the nails. Data recording was performed by MATLAB 7.8 (MathWorks Inc., USA).

Procedure

Participants took part in two sessions on subsequent days, one practice session and one experimental session. The practice session familiarized participants with the tasks, thus removing the need for training blocks in the experimental session. In the practice session, both participants simultaneously performed 10 pointing and 10 grasping movements before performing four blocks of 30 trials of both movements randomly intermixed. The participant on the left used the left hand while the participant on the right used the right hand. After two blocks, participants switched seating positions and used the other hand. Both participants then simultaneously performed four blocks of 30 trials of the combined task bimanually, using the left hand for the movement task and the right hand for the search task. After two blocks, participants again switched seating positions.

The experimental session consisted of 12 blocks of 60 trials. In the first four blocks, one participant was seated on the left and performed the movement task, while the other was seated on the right and performed the search task. After the fourth block, participants switched seating positions and performed the other task for another four blocks. In the last four blocks, participants performed their initial task again.

Experimental trials started with a fixation cross shown for 300 ms. Then, the movement cue was presented for 1000 ms. Next, a fixation cross was shown for a randomly chosen duration of 200–400 ms, followed by the search display presented for 100 ms. Another fixation cross was presented while the searcher indicated whether a target was present or absent in the search display by pressing one of two buttons on the response box. The searcher was asked to respond as fast as possible while maintaining an accuracy of over 85%. Button assignment (left or right button for target presence) was counterbalanced across participant pairs. The fixation cross remained on screen until a response was made or 1800 ms after search display offset. After another 100 ms, the Go cue was presented for 300 ms indicating the movement goal object. The performer was instructed to execute the prepared movement as fast as possible with cue onset. Correctness of the movement was registered by the experimenter seated 50 cm behind and 50 cm to the left of the performer. If the movement was not initiated within 1800 ms after movement cue onset (as registered by the release of the starting position button plate), a text display was shown ("no movement") and the trial ended. At the end of each block, a

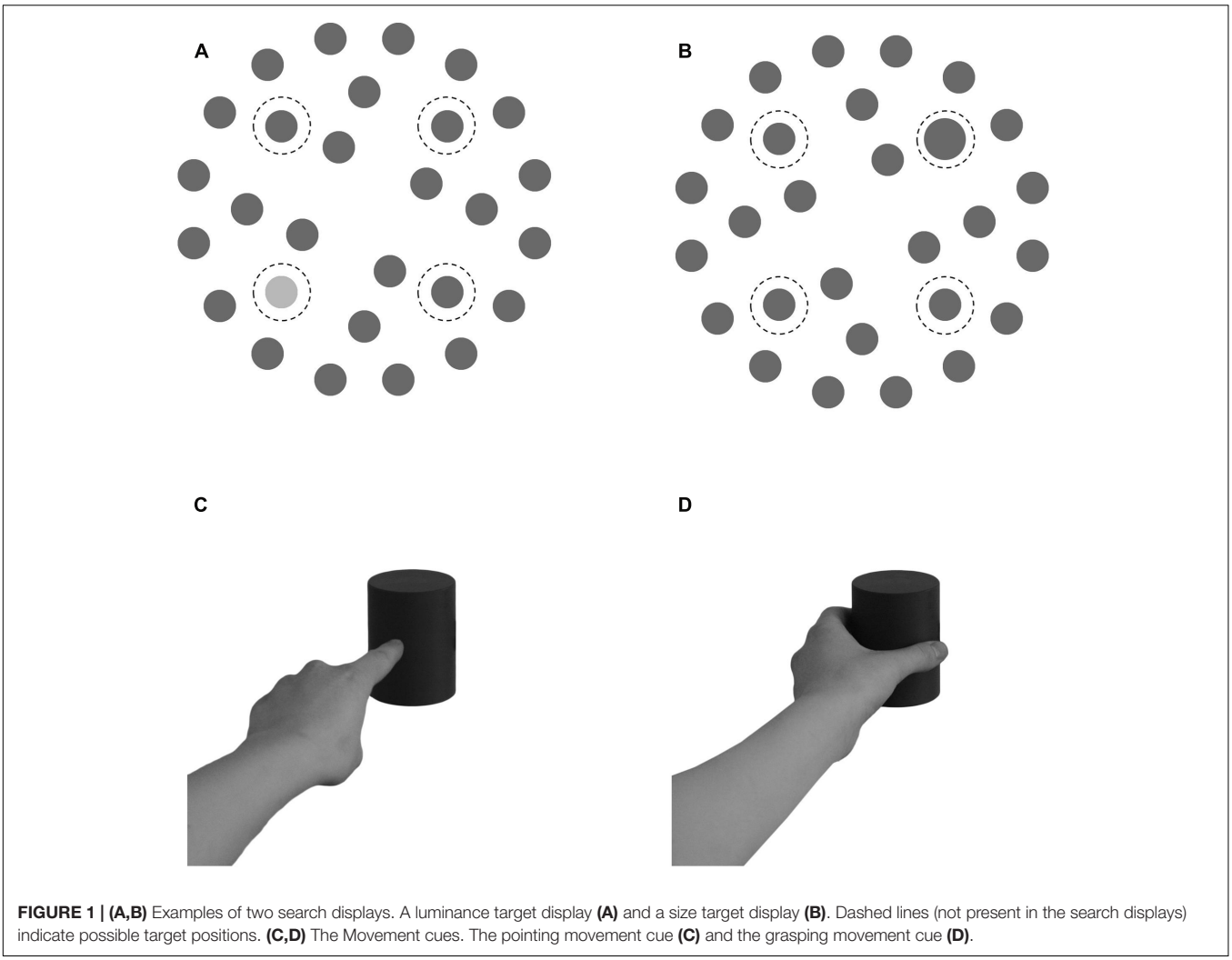


FIGURE 1 | (A,B) Examples of two search displays. A luminance target display **(A)** and a size target display **(B)**. Dashed lines (not present in the search displays) indicate possible target positions. **(C,D)** The Movement cues. The pointing movement cue **(C)** and the grasping movement cue **(D)**.

feedback screen showed the searcher's mean RT and accuracy in the search task together with the performer's mean time of movement onset, movement duration, and movement accuracy. Participants were asked not to talk during a block and to pause between blocks when necessary. A new combination of randomly selected movement objects was set up for each experimental block.

The search target type (luminance or size) remained constant for two subsequent blocks, with the order counterbalanced across participant pairs.

Data Analysis

Search Task

For RT analysis in the search task, mean RTs were computed for each participant and each block separately. Outlier trials (± 2 standard deviations from participants' mean RT in the corresponding block) as well as trials with inaccurate or no responses were excluded from further analysis. For the analysis of response accuracy in the search task, only outlier trials were excluded. To investigate whether the performer's movement task affected the searcher's RTs and accuracies in the search task, hierarchical linear mixed models (HLM) were used to predict the searcher's performance. Thereby, in addition to controlling for within-subject data dependencies as in repeated measure ANOVAs, we considered the dependent data structure of participants nested in pairs who switched tasks during the experiment. The HLMs were based on the experimental factors in every single trial, rather than on individual participants' mean data for one experimental factor or factor combination as in ANOVA procedures. Using HLMs had two advantages: Higher statistical power compared to ANOVA procedures and controlling for dependencies on multiple data levels. Pairs of participants were modeled on the highest analysis level, individual participants on a second level and the three experimental parts separated by participants switching tasks (first, second, and third four blocks) with two subsequent blocks of each target type on the lowest level. Random intercepts were included in the model to account for dependencies within these data units. Target type (luminance vs. size), trial type (target absent vs. target present), movement type (grasp vs. point), and experimental part (first vs. second vs. third) were introduced

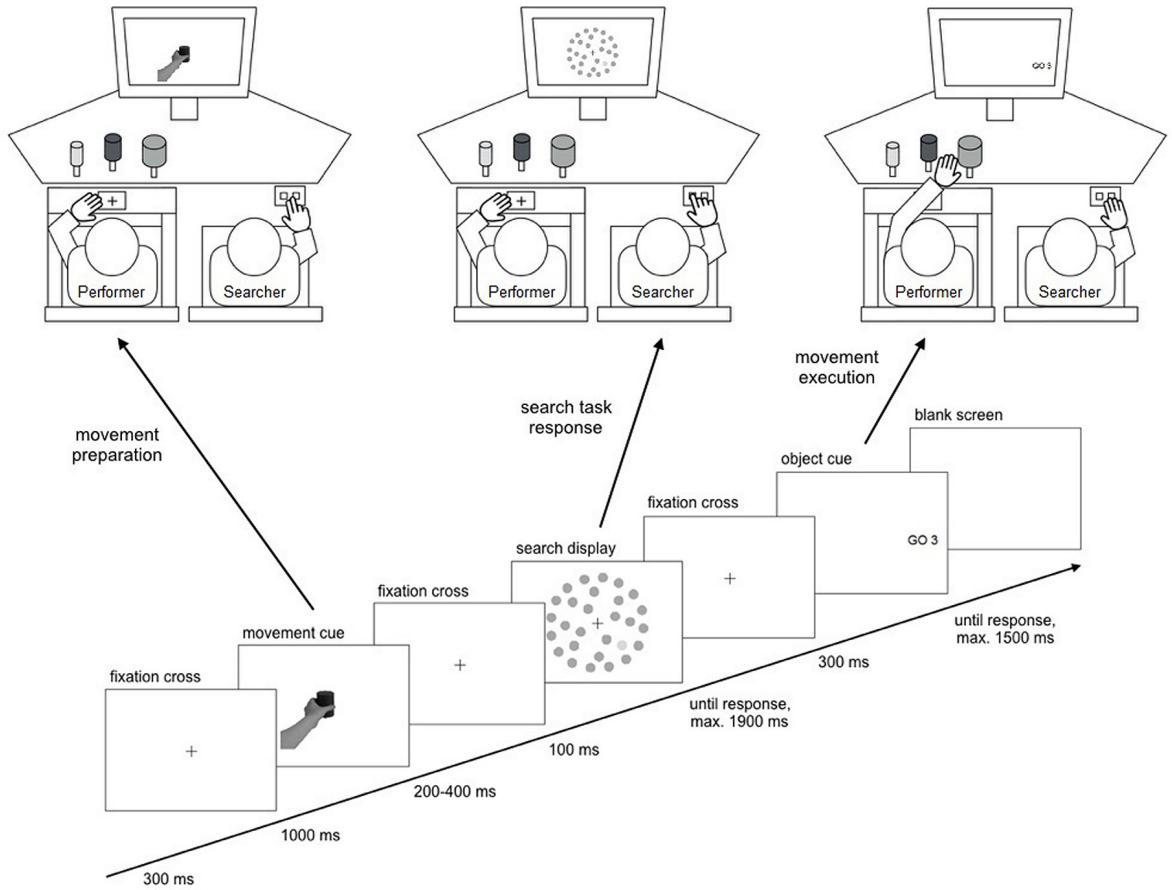


FIGURE 2 | Schematic of the experimental setup and the trial sequence. Performers first received a movement cue instructing them to prepare either a grasping or a pointing movement. Then the searcher performed a visual search for a size or a luminance target. A Go cue then indicated one of three goal objects for the performer's movement execution.

as fixed effects, which can be interpreted similarly to within-subject factors of an ANOVA procedure. All possible two-way interactions between target type, trial type and movement type and the three-way interaction were specified. Additionally, interactions between experimental part and the aforementioned effects were specified to investigate whether the modulation of the searcher's search performance by the performer's movement task differed depending on which task participants performed initially, and whether the modulation changed between the first and third part of the experiment. Because participants searched for each target type twice in pairs of subsequent blocks, a fixed effect was included to account for learning effects from the first to the second block of each block pair. Significant effects were followed up by simple main effect pairwise comparisons based on estimated marginal means, corrected for multiple comparisons via Bonferroni adjustments of the critical *p*-values.

Movement Task

Positional data from the sensors on the performer's thumb and index finger was used to analyze movement performance. A fourth order low-pass Butterworth filter with a cut-off frequency of 20 Hz was applied to smooth sensor velocity data. Two dependent variables were computed to reflect the beginning and end of each movement: time of movement onset (MO) and mean movement velocity (MV). MV was chosen rather than movement duration to measure efficiency of movement execution as the distance between starting position and the goal objects was different for each participant depending on the comfortable reaching distance determined at the beginning of the experiment. MO was calculated as the time from Go cue presentation to the point when the velocity of the performer's index finger sensor first exceeded 10 cm/s. To calculate MV, the point in time after MO when the performer's index finger was resting on the goal object was identified. All data samples of a trial were considered where the index finger sensor was further away than 20 cm from its position at MO. The sample in that data range where the velocity of the index finger sensor was at its minimum was considered as the point in time when the performer's index finger rested on the goal object. MV was calculated as the time from MO until the resting

point divided by the distance (displacement) between the index finger sensor at that resting point and its position at MO.

Out of the 65 participants included in the analysis of search performance, movement data recording failed for 12 participants, probably due to technical error during sensor application while switching tasks. This resulted in availability of half or less of all trials of these participants. They were therefore excluded from movement performance analysis. MO and MV data were again analyzed using HLM models. Only trials where the correct movement was performed were included. Outlier trials were excluded according to the same criterion as in search performance analyses and factors were specified analogous to search performance analyses (see above).

RESULTS

Search Task

Response Times

Response times differed significantly depending on the target type, $F(1,95.3) = 176$, $p < 0.001$, with longer RTs for size target detection [estimated marginal mean (M) = 516 ms, standard error of the mean (SEM) = 12.4 ms] than for luminance target detection (M = 451 ms, SEM = 12.4 ms). RTs also differed significantly depending on the trial type, $F(1,20660) = 547$, $p < 0.001$, with longer RTs for target absent trials (M = 499 ms, SEM = 12.2 ms) than for target present trials (M = 469 ms, SEM = 12.2 ms). Importantly, there was a significant interaction between target type and movement type, $F(1,20655) = 5.71$, $p = 0.017$. Participants' mean RTs were calculated to illustrate this interaction, depicted in **Figure 3A**. Pairwise comparisons showed that RTs in luminance target trials were longer for pointing (M = 453 ms, SEM = 12.4 ms) than for grasping movements (M = 449.3 ms, SEM = 12.4 ms), MD = 4.14 ms, SEM = 1.74 ms, df = 20655, $p = 0.018$, while RTs in size target trials were not significantly different for pointing (M = 515 ms, SEM = 12.4 ms) and grasping movements (M = 517.1 ms, SEM = 12.4 ms), $MD = -1.88$ ms, SEM = 1.82 ms, df = 20656, $p = 0.301$. There was a significant interaction between target type and trial type, $F(1,20663) = 7.93$, $p = 0.005$. Pairwise comparisons based on estimated marginal means showed that this was due to a larger RT difference between target absent and target present trials for size targets (MD = 33.1 ms, 95% CI [29.5 ms, 36.7 ms]) than for luminance targets (MD = 26.0 ms, 95% CI [22.6 ms, 29.4 ms]).

Response times decreased between the first and the second of two subsequent blocks in which participants searched for one target type, MD = 23.7 ms, SEM = 1.26 ms, $F(1,20655) = 353$, $p < 0.001$. RTs also differed significantly depending on the experimental part, $F(2,39.2) = 13.7$, $p < 0.001$. Pairwise comparisons showed that overall, RTs decreased from experimental parts 1 to 3 and were shorter in experimental part 3 than in part 2, while there was no difference between experimental parts 1 and 2. There was also a significant interaction between experimental part, target type and trial type, $F(2, 20663) = 7.80$, $p < 0.001$, indicating that there was no RT difference between experimental parts 2 and 3 for luminance target present trials (see **Table 1** for follow-up pairwise comparisons).

Search Accuracy

Accuracies of search responses differed significantly depending on the target type, $F(1,95.1) = 103$, $p < 0.001$, with higher accuracies for luminance target detection (M = 96.9%, SEM = 0.84%) than for size target detection (M = 89.3%, SEM = 0.84%). Response accuracies also differed significantly depending on the trial type, $F(1,22206) = 232$, $p < 0.001$, with higher accuracies in target absent trials (M = 95.5%, SEM = 0.76%) than in target present trials (M = 90.7%, SEM = 0.76%). Again, there was a significant interaction between target type and movement type, $F(1,22206) = 6.19$, $p = 0.013$. Participants' mean search accuracies were calculated to illustrate this interaction, depicted in **Figure 3B**. Pairwise comparisons showed that accuracies in size target trials were higher for pointing (M = 89.9%, SEM = 0.87%) than for grasping movements (M = 88.7%, SEM = 0.87%), $MD = -1.24$%, SEM = 0.45%, df = 22205, $p = 0.006$, while accuracies in luminance target trials did not differ between pointing (M = 96.7%, SEM = 0.87%) and grasping movements (M = 97.1%, SEM = 0.87%), $MD = -0.36$%, SEM = 0.46%, df = 22206, $p = 0.431$. There was also a significant interaction between target type and trial type, $F(1,22207) = 97.9$, $p < 0.001$. Pairwise comparisons showed that this was due to a larger accuracy difference between target absent and target present trials for size targets (MD = 8.07%, 95% CI [7.18%, 8.96%]) than for luminance targets (MD = 1.71%, 95% CI [0.82%, 2.60%]).

Accuracies of search responses increased between the first and the second of two subsequent blocks in which participants searched for one target type [MD = 2.19%, SEM = 0.32%, $F(1,22202) = 46.4$, $p < 0.001$]. There was a significant interaction between experimental part and trial type, $F(2,22207) = 4.30$, $p = 0.014$. Pairwise comparisons showed that this was due to significantly higher accuracies of responses in experimental part 3 than in part 1 in target absent trials while response accuracies did not differ between any two experimental parts in target present trials (**Table 1**).

Movement Task

Movement Onset

There was a significant interaction between target type and trial type $F(1,161456) = 4.65$, $p = 0.018$, reflecting that MOs were longer in size target present trials (M = 363 ms, SEM = 10.6 ms) than in size target absent trials (M = 361 ms, SEM = 10.6 ms), while MOs were shorter in luminance target present trials (M = 359 ms, SEM = 10.6 ms) than in luminance target absent trials (M = 361 ms, SEM = 10.6 ms). However, in pairwise comparisons, neither MO difference was found to be significant (size targets: $MD = -1.80$ ms, SEM = 1.25 ms, df = 16145, $p = 0.149$; luminance targets: $MD = -1.78$ ms, SEM = 1.23 ms, df = 16146, $p = 0.146$).

Movement onsets decreased between the first and the second of two subsequent blocks in which participants searched

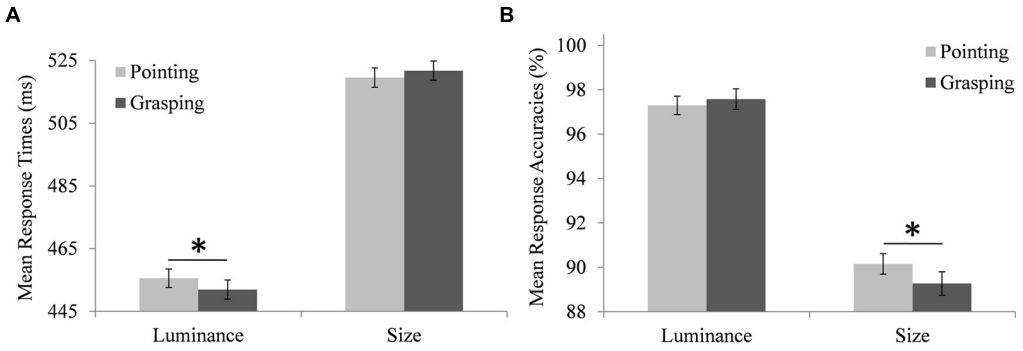


FIGURE 3 | Modulation of the searchers' performance in the search task by the performer's prepared movement. (A) Searchers' mean response times to luminance and size targets when the performer prepared a pointing movement (light bars) or a grasping movement (dark bars). **(B)** Mean accuracies of searchers' responses to luminance and size targets when the performer prepared a pointing movement (light bars) or a grasping movement (dark bars). Note that the depicted means are based on individual searchers' aggregated data, while the employed hierarchical linear models used individual trial data. Error bars represent standard errors adjusted for within-subject designs, calculated according to the procedure described in Cousineau (2005). *Indicate significant differences in pairwise comparisons ($p < 0.05$, corrected for multiple comparisons via Boferroni adjustments).

for one target type [$MD = 9.96$ ms, $SEM = 0.88$ ms, $F(1,16154) = 129.3$, $p < 0.001$]. MOs differed significantly between experimental parts, $F(2,27.7) = 4.65$, $p = 0.018$, and there was a significant interaction of experimental part and movement type, $F(2,16146) = 4.93$, $p = 0.007$. Pairwise comparisons showed that this was due to a higher decrease in MOs of grasping movements between experimental parts 1 and 3 compared to pointing movements (**Table 1**).

Movement Velocity

There was a significant interaction of target type, trial type and movement type, $F(1,16181) = 5.02$, $p = 0.025$. Participants' mean MVs were calculated to illustrate this three-way interaction, depicted in **Figure 4**. Pairwise comparisons showed that the interaction reflected lower MVs of grasping movements in size target present trials compared to size target absent trials ($MD = -0.67$ cm/s, $SEM = 0.25$ ms, $df = 16181$, $p = 0.008$), while the MV difference between pointing movements in luminance target present trials and luminance target absent trials was not significant ($MD = -0.36$ cm/s, $SEM = 0.26$ ms, $df = 16181$, $p = 0.168$). There was no effect of trial type on MVs of pointing movements in size target trials ($MD = -0.01$ cm/s, $SEM = 0.25$ ms, $df = 16181$, $p = 0.976$) and no effect of trial type on MVs of grasping movements in luminance target trials ($MD = -0.11$ cm/s, $SEM = 0.25$ ms, $df = 16180$, $p = 0.664$).

Movement velocities increased between the first and the second of two subsequent blocks in which participants searched for one target type [$MD = 1.50$ cm/s, $SEM = 1.23$ cm/s, $F(1,16186) = 139.3$, $p < 0.001$]. MVs differed significantly between experimental parts, $F(2,29.1) = 3.41$, $p = 0.047$, and there was a significant interaction of experimental part and movement type, $F(2,16181) = 9.81$, $p < 0.001$. Pairwise comparisons showed that this was likely due to increased MVs of pointing movements in experimental part 3 compared to part 1, while MVs of grasping movement did not increase (**Table 1**).

DISCUSSION

The present study examined whether a partner's action planning modulates an agent's perception in a joint action task. A paradigm previously used to demonstrate that action planning can affect perceptual processing of action-relevant dimensions in individual agents (Wykowska et al., 2009) was adapted so that two participants sitting side-by-side could perform the task conjointly. While one participant (the "performer") prepared to perform a pointing or grasping movement, the other participant ("the searcher") searched for either a luminance or a size target on a computer screen.

Results showed that the movement the performer was preparing modulated the searcher's perceptual performance. In luminance target trials, RTs were longer in the search task when the performer prepared a pointing movement compared to a grasping movement. Accuracy of search responses also indicated a modulation of the searcher's performance by the performer's prepared movement, mirroring RT results: Responses to size targets were less accurate when the performer prepared a grasping movement compared to a pointing movement. Similarly, the search task influenced the performer's movement execution, although this effect was less pronounced. When the searcher was searching for a size target, the performer executed grasping movements with lower velocity when a target was present in the display than when it was absent.

Importantly, the modulation of the searcher's performance by the performer's movement was observed before the actual execution of the movement. The searcher processed the search targets differently depending on the movement the performer was preparing to execute subsequently. As there was no perceptual difference in the search task for trials requiring a subsequent pointing or grasping movement, the finding that the performer's prepared movement modulated the searcher's perceptual processing indicates that the searcher represented features relevant to the performer's movement in addition to the features relevant to the own visual search task. Hence the

TABLE 1 | Pairwise comparisons for significant fixed effects of the hierarchical linear models including the experimental part.

Dependent variable	Movement	Target type	Trial type	Part	*MD*	*SEM*	*df*	*p*
Search response time (ms)		Luminance	Target present	1–2	2.8	16.1	46.6	>0.999
				1–3	28.9	10.0	81.7	0.014
				2–3	26.1	16.1	46.6	0.334
			Target absent	1–2	−23.0	16.1	46.6	0.478
				1–3	27.7	10.0	81.7	0.020
				2–3	50.7	16.1	46.6	0.008
		Size	Target present	1–2	−7.8	16.1	47.1	>0.999
				1–3	35.3	10.0	83.7	0.002
				2–3	43.1	16.1	47.1	0.031
			Target absent	1–2	−14.4	16.1	46.7	>0.999
				1–3	56.6	10.0	82.4	<0.001
				2–3	71.0	16.1	46.6	<0.001
Search accuracy (%)								
			Target present	1–2	0.05	1.19	55.5	>0.999
				1–3	−0.93	1.07	56.2	>0.999
				2–3	−0.97	1.19	55.5	>0.999
			Target absent	1–2	−1.70	1.19	55.6	0.477
				1–3	−3.09	1.08	56.5	0.017
				2–3	−1.39	1.19	55.6	0.745
Movement onset (ms)								
	Pointing			1–2	13.4	19.3	30.8	>0.999
				1–3	27.7	10.2	19.9	0.040
				2–3	14.4	19.0	29.6	>0.999
	Grasping			1–2	14.9	19.3	30.8	>0.999
				1–3	34.1	10.2	19.9	0.010
				2–3	19.3	19.0	29.6	0.961
Movement velocity (cm/s)								
	Pointing			1–2	−6.34	3.17	29.9	>0.163
				1–3	−4.28	1.67	22.2	0.053
				2–3	2.06	3.13	28.7	>0.999
	Grasping			1–2	−4.94	3.17	29.9	0.386
				1–3	−3.50	1.67	22.2	0.145
				2–3	1.45	3.13	28.7	>0.999

Movement: Movement type. Part: Experimental part (blocks 1–4 vs. blocks 5–8 vs. blocks 9–12). MD: Mean difference based on estimated marginal means. Critical p-values were corrected for multiple comparisons via Bonferroni adjustments.

searcher represented the performer's movement, likely similar to an own movement and relying on the common-coding format of perception and action (Prinz, 1990; Hommel et al., 2001; Hommel, 2009).

Both search RT and accuracy results suggest that representing the features of the partner's movement impaired the searcher's perception rather than facilitating it. In trials when the search target dimension was congruent to the dimension relevant to the partner's movement, search RTs were longer and accuracy was lower compared to incongruent trials. This is not in line with results observed in the single agent version of the paradigm, which reported facilitation of perception by congruent action planning (Wykowska et al., 2009, 2012; Wykowska and Schubö, 2012). Instead, the present results match previous joint action research indicating an interfering influence of a partner's task (e.g., Sebanz et al., 2003).

To explain the present results, one may argue as follows: A prepared movement of the performer activated the action-relevant feature dimension also in the searcher. In incongruent trials, the representation of the features relevant to the performer's movement did not impose an additional load on the searcher's perceptual system, as this representation included different perceptual dimensions than the one required to detect the target. In congruent trials, however, the representation of the features relevant to the performer's movement included

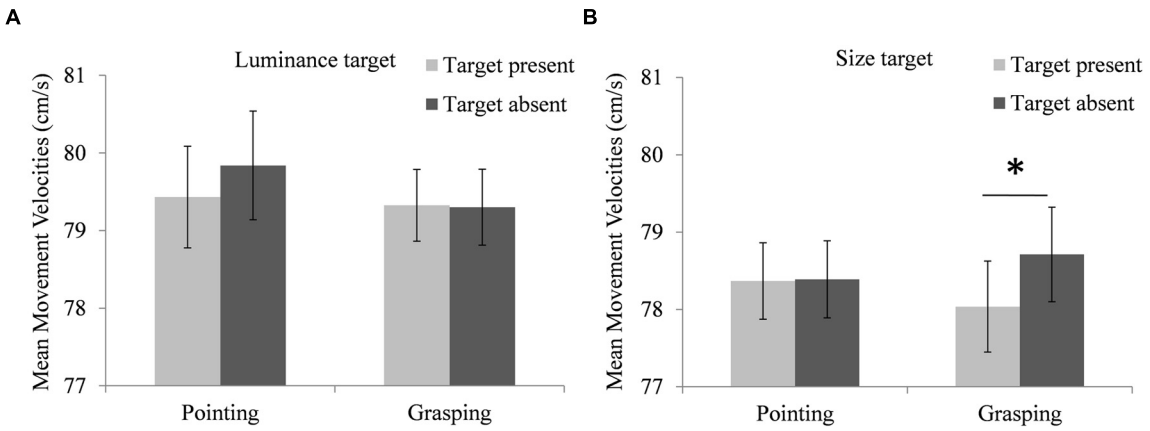


FIGURE 4 | Modulation of the performers' movement task performance by the searcher's target dimension. (A) Performers' mean velocities of pointing and grasping movements in luminance target present trials (light bars) and luminance target absent trials (dark bars). **(B)** Performers' mean velocities of pointing and grasping movements in size target present trials (light bars) and size target absent trials (dark bars). Note that the depicted means are based on individual performers' aggregated data, while the employed hierarchical linear models used individual trial data. Error bars represent standard errors adjusted for within-subject designs, calculated according to the procedure described in Cousineau (2005). *Indicate significant differences in pairwise comparisons ($p < 0.05$, corrected for multiple comparisons via Boferroni adjustments).

the perceptual dimension that was required to detect the target. When the performer prepared a pointing movement in luminance target trials, for instance, the searcher needed to discern whether the activation of the luminance dimension resulted from the detection of a target in the search display (i.e., the own task), or from the representation of the pointing movement (i.e., the performer's task). The cost of this additional process may explain the prolonged RTs in these trials.

For size targets, RTs descriptively followed the same congruency pattern, with longer RTs when the performer prepared a grasping movement compared to a pointing movement, but the difference was not significant. A ceiling effect due to a generally higher difficulty of size target detection, as evidenced by longer RTs and lower accuracies for size than for luminance targets, may explain this. For size targets, the difference between movement types manifested in lower accuracies of search responses when the performer prepared a grasping movement compared to a pointing movement. This can also be considered an indication of the additional load on the searcher's perceptual system when the features relevant to the performer's movement were required for detecting the target. Again, accuracies of luminance target detection followed the same congruency pattern, with lower accuracies of search responses for pointing compared to grasping movements, but the difference was not significant. Here, the generally lower difficulty of detecting luminance targets compared to size targets might have reduced an impairing influence of the representation of features relevant to the performer's pointing movement on response accuracies of luminance target detection.

The PCM of motorvisual priming (Thomaschke et al., 2012) was suggested to explain action planning effects on perception. According to PCM, the direction of modulatory effects of planned actions on perceptual processes depends on whether the action can be fully specified in advance or whether it requires the online adjusting of open parameters. PCM assumes that action planning temporarily binds representations of features of the planned action. Feature dimensions bound in this process are less accessible to other cognitive processes, including perception. In the present paradigm, the searcher may have bound features relevant to the performer's movement although they were not directly relevant to the search task, simply as a consequence of the joint action context. Previous results have shown that agents tend to form representations of their partner's part in joint action tasks (Vesper et al., 2010). In the present paradigm, such feature binding led to an impairment of performance in the search task in congruent trials. In incongruent trials, however, the bound features did not match the dimension relevant to search target detection, thus binding did not impair perceptual performance.

According to PCM, perceptual facilitation by action planning is observed when an action cannot be not fully specified in advance but requires online adjusting of open parameters. This was the case in the single agent version of the present paradigm (Wykowska et al., 2009, 2012; Wykowska and Schubö, 2012) where participants had to wait for the Go cue to identify the goal object of the prepared movement. Only then could the movement execution be adjusted to the location and size of the movement goal. In contrast, the searcher did not represent the performer's movement as a partially unspecified action with open parameters in the present paradigm. The key difference between the joint and single agent task is that the searcher does not execute the planned movement in the joint action task, hence adjusting open movement parameters is not required. Instead, the searcher has to suppress any tendency to execute the movement. Thus, the searcher never switches from action planning to the movement control system, which, according to PCM, causes facilitation of perception by action planning.

Although our main research question focused on the searcher's performance, we also investigated whether the congruency between the searcher's target dimension and the dimension relevant to the performer's movement influenced movement performance. When the searcher searched for size targets, velocities of grasping movements were lower in target present compared to target absent trials. Similarly, velocities of pointing movements were numerically lower in luminance target present trials compared to target absent trials. This finding can be interpreted in a similar way as the impaired search performance in congruent compared to incongruent trials. During movement preparation, the performer attended the search display to execute the movement as soon as the Go cue appeared on the screen. Therefore, the performer perceptually processed the target at least in some trials. In size target present trials, processing the target activated the size dimension in the performer's perceptual system. The additional load on this system then impaired grasping efficiency in these trials. In target absent trials, no size information was available, leaving more resources for the grasping movement.

In general, variability in movements was larger than variability in search responses. This may have been a consequence of movement types being randomly intermixed within blocks, making it harder for the performer to switch between movement types, while the search target remained the same for two subsequent blocks. Interestingly, performers showed a general tendency to adapt their movement execution to searchers' performance. Correlation analyses showed that trials with longer search RTs also had later MOs, $r(16320) = 0.12$, $p < 0.001$, and movements were executed with lower velocities when search RTs were longer, $r(16320) = -0.22$, $p < 0.001$.

Performance in both tasks improved between subsequent blocks: Search RTs decreased and accuracies increased, while MOs were earlier and movements were executed with higher velocities in the second of two subsequent blocks. This indicates a short-term learning effect in both the searcher and the performer. Performance also generally increased when participants returned to the same task (i.e., from experimental parts 1–3), pointing toward a benefit of prior task experience. Performance did not differ for participants who performed the search or the movement task first. Importantly, the observed action-perception effects were also not different, and neither differed between experimental parts 1 and 3. This suggests that task order had no impact on the observed modulation of perceptual processing by action planning across participants.

Why do agents tend to represent a partner's task although this is not necessary or even detrimental to performing the own task? For instance, how does the representation of the performer's planned movement benefit the searcher in the present paradigm? Predictive coding accounts of human cognition postulate that the brain's higher-level cortical systems predict the input to lower-level systems. Perception constitutes the lowest level of information in this multidirectional hierarchical system. Comparisons to sensory feedback cause higher-level systems to adapt to reduce the size of prediction errors (Clark, 2013). Likewise, agents act in such a way that the resulting sensory inputs match the predicted sensory outcomes as closely as possible (Friston, 2010). Consistent with this view, we assume that knowing which movement a partner is planning reduces the prediction error in the joint task. Thus, by representing the partner's movement similar to an own movement in the common-coding format of perception and action (Prinz, 1997; Hommel et al., 2001; Hommel, 2009) the agent maximizes the predictability of joint action outcomes. This gain in predictability appears to outweigh the potential additional cost of representing the partner's task. Together with previous findings on action simulation (Rizzolatti and Craighero, 2004; Fadiga et al., 2005; Newman-Norlund et al., 2008; Avenanti et al., 2013), the present results suggest that the same systems are utilized to establish this cross-brain predictive coding system that the agent usually employs to represent an own action. Predictive coding may thus also operate across brains to provide agents with information about a partner's actions when they coordinate to reach a common goal.

ETHICS STATEMENT

All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments. Informed consent was obtained from all individual participants included in the study.

AUTHOR CONTRIBUTIONS

AS and DD designed the experiment; DD conducted data analyses; CV suggested further analyses of movement parameters; AS, CV, and DD discussed results and contributed to the manuscript. All authors have approved the current version of the article.

FUNDING

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; SCHU 1330/5-1 and SFB/TRR 135, TP B3).

REFERENCES

Avenanti, A., Candidi, M., and Urgesi, C. (2013). Vicarious motor activation during action perception: beyond correlational evidence. *Front. Hum. Neurosci.* 7:185. doi: 10.3389/fnhum.2013.00185

Baess, P., and Prinz, W. (2015). My partner is also on my mind: social context modulates the N1 response. *Exp. Brain Res.* 233, 105–113. doi: 10.1007/s00221-014-4092-9

Bekkering, H., de Bruijn, E. R. A., Cuijpers, R. H., Newman-Norlund, R., van Schie, H. T., and Meulenbroek, R. (2009). Joint action: neurocognitive mechanisms

supporting human interaction. *Top. Cogn. Sci.* 2, 340–352. doi: 10.1111/j.1756-8765.2009.01023.x

Bekkering, H., and Neggers, S. F. W. (2002). Visual search is modulated by action intentions. *Psychol. Sci.* 13, 371–374. doi: 10.1111/j.0956-7976.2002.00466.x

Brass, M., Bekkering, H., and Prinz, W. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychol.* 106, 3–22. doi: 10.1016/S0001-6918(00)00024-X

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Cohen, R. G., and Rosenbaum, D. A. (2004). Where grasps are made reveals how grasps are planned: generation and recall of motor plans. *Exp. Brain Res.* 157, 486–495. doi: 10.1007/s00221-004-1862-9

Cousineau, D. (2005). Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's Method. *Tutor. Quant. Methods Psychol.* 1, 42–45. doi: 10.3758/s13423-012-0230-1

Craighero, L., Fadiga, L., Rizzolatti, G., and Umilta, C. (1999). Action for perception: a motor-visual attentional effect. *J. Exp. Psychol. Hum.* 25, 1673–1692. doi: 10.1037/0096-1523.25.6.1673

Cross, E. S., Hamilton, A. F., de, C., and Grafton, S. T. (2006). Building a motor simulation de novo: observation of dance by dancers. *Neuroimage* 31, 1257–1267. doi: 10.1016/j.neuroimage.2006.01.033

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Exp. Brain Res.* 91, 176–180. doi: 10.1007/BF00230027

Dötsch, D., and Schubö, A. (2015). Social categorization and cooperation in motor joint action: evidence for a joint end-state comfort. *Exp. Brain Res.* 233, 2323–2334. doi: 10.1007/s00221-015-4301-1

Fadiga, L., Craighero, L., and Olivier, E. (2005). Human motor cortex excitability during the perception of others' action. *Curr. Opin. Neurobiol.* 15, 213–218. doi: 10.1016/j.conb.2005.03.013

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Glover, S. (2004). Separate visual representations in the planning and control of action. *Behav. Brain Sci.* 27, 3–24. doi: 10.1017/S0140525X04000020

Güldenpenning, I., Steinke, A., Koester, D., and Schack, T. (2013). Athletes and novices are differently capable to recognize feint and non-feint actions. *Exp. Brain Res.* 230, 333–343. doi: 10.1007/s00221-013-3658-2

Hommel, B. (2009). Action control according to TEC (theory of event coding). *Psychol. Res.* 73, 512–526. doi: 10.1007/s00426-009-0234-2

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–937. doi: 10.3389/2Ffpsyg.2015.01318

Memelink, J., and Hommel, B. (2013). Intentional weighting: a basic principle in cognitive control. *Psychol. Res.* 77, 249–259. doi: 10.1007/s00426-012-0435-y

Ménoret, M., Curie, A., des Portes, V., Nazir, T. A., and Paulignan, Y. (2013). Motor resonance facilitates movement execution: an ERP and kinematic study. *Front. Hum. Neurosci.* 7:64. doi: 10.3389/fnhum.2013.00646

Müsseler, J., and Hommel, B. (1997a). Blindness to response-compatible stimuli. *J. Exp. Psychol. Hum.* 23, 861–872. doi: 10.1037/0096-1523.23.3.861

Müsseler, J., and Hommel, B. (1997b). Detecting and identifying response-compatible stimuli. *Psychon. Bull. Rev.* 4, 125–129. doi: 10.3758/BF03210785

Newman-Norlund, R., Bosga, J., Meulenbroek, R., and Bekkering, H. (2008). Anatomical substrates of cooperative joint-action in a continuous motor task: virtual lifting and balancing. *Neuroimage* 41, 169–177. doi: 10.1016/j.neuroimage.2008.02.026

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Prinz, W. (1990). "A common-coding approach to perception and action," in *Relationships Between Perception and Action*, eds O. Neumann and W. Prinz (Berlin: Springer), 167–201.

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154. doi: 10.1080/713752551

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci.* 10, 70–76. doi: 10.1016/j.tics.2005.12.009

Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition* 88, 11–21. doi: 10.1016/S0010-0277(03)00043-X

Simon, J. R., and Rudell, A. P. (1967). Auditory S-R compatibility: the effect of an irrelevant cue on information processing. *J. Appl. Psychol.* 51, 300–304. doi: 10.1037/h0020586

Springer, A., Parkinson, J., and Prinz, W. (2013). Action simulation: time course and representational mechanisms. *Front. Psychol.* 4:387. doi: 10.3389/fpsyg.2013.00387

Symes, E., Tucker, M., Ellis, R., Vainio, L., and Ottoboni, G. (2008). Grasp preparation improves change detection for congruent objects. *J. Exp. Psychol. Hum.* 34, 854–871. doi: 10.1037/0096-1523.34.4.854

Thomaschke, R. (2012). Investigating ideomotor cognition with motorvisual priming paradigms: key findings, methodological challenges, and future directions. *Front. Psychol.* 3:512. doi: 10.3389/fpsyg.2012.00519

Thomaschke, R., Hopkins, B., and Miall, R. C. (2012). The planning and control model (PCM) of motorvisual priming: reconciling motorvisual impairment and facilitation effects. *Psychol. Rev.* 119, 388–407. doi: 10.1037/a0027453

Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010). A minimal architecture for joint action. *Neural Netw.* 23, 998–1003. doi: 10.1016/j.neunet.2010.06.002

Wykowska, A., Hommel, B., and Schubö, A. (2012). Imaging when acting: picture but not word cues induce action-related biases of visual attention. *Front. Psychol.* 3:388. doi: 10.3389/fpsyg.2012.00388x

Wykowska, A., and Schubö, A. (2012). Action intentions modulate allocation of visual attention: electrophysiological evidence. *Front. Psychol.* 3:379. doi: 10.3389/fpsyg.2012.00379

Wykowska, A., Schubö, A., and Hommel, B. (2009). How you move is what you see: action planning biases selection in visual search. *J. Exp. Psychol. Hum.* 35, 1755–1769. doi: 10.1037/a0016798

Conflict of Interest Statement: 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.

Copyright © 2017 Dötsch, Vesper and Schubö. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 November 2016
doi: 10.3389/fpsyg.2016.01845

Task-Irrelevant Expectation Violations in Sequential Manual Actions: Evidence for a "Check-after-Surprise" Mode of Visual Attention and Eye-Hand Decoupling

*Rebecca M. Foerster**

Neuro-cognitive Psychology, Department of Psychology & Cluster of Excellence Cognitive Interaction Technology 'CITEC', Bielefeld University, Bielefeld, Germany

When performing sequential manual actions (e.g., cooking), visual information is prioritized according to the task determining where and when to attend, look, and act. In well-practiced sequential actions, long-term memory (LTM)-based expectations specify which action targets might be found where and when. We have previously demonstrated (Foerster and Schneider, 2015b) that violations of such expectations that are task-relevant (e.g., target location change) cause a regression from a memory-based mode of attentional selection to visual search. How might task-irrelevant expectation violations in such well-practiced sequential manual actions modify attentional selection? This question was investigated by a computerized version of the number-connection test. Participants clicked on nine spatially distributed numbered target circles in ascending order while eye movements were recorded as proxy for covert attention. Target's visual features and locations stayed constant for 65 prechange-trials, allowing practicing the manual action sequence. Consecutively, a task-irrelevant expectation violation occurred and stayed for 20 change-trials. Specifically, action target number 4 appeared in a different font. In 15 reversion-trials, number 4 returned to the original font. During the first task-irrelevant change trial, manual clicking was slower and eye scanpaths were larger and contained more fixations. The additional fixations were mainly checking fixations on the changed target while acting on later targets. Whereas the eyes repeatedly revisited the task-irrelevant change, cursor-paths remained completely unaffected. Effects lasted for 2–3 change trials and did not reappear during reversion. In conclusion, an unexpected task-irrelevant change on a task-defining feature of a well-practiced manual sequence leads to eye-hand decoupling and a "check-after-surprise" mode of attentional selection.

Keywords: eye movements, attention, expectation violation, surprise, manual action sequence, sensorimotor learning, eye-hand coupling

OPEN ACCESS

Edited by:
Anna Thorwart,
University of Marburg, Germany

Reviewed by:
Ulrich Ansorge,
University of Vienna, Austria
Robert Gaschler,
FernUniversität in Hagen, Germany

***Correspondence:**
Rebecca M. Foerster
rebecca.foerster@uni-bielefeld.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *29 September 2016*
Accepted: *07 November 2016*
Published: *23 November 2016*

Citation:
Foerster RM (2016) Task-Irrelevant Expectation Violations in Sequential Manual Actions: Evidence for a "Check-after-Surprise" Mode of Visual Attention and Eye-Hand Decoupling. Front. Psychol. 7:1845. doi: 10.3389/fpsyg.2016.01845

INTRODUCTION

When performing a manual action sequence in an unfamiliar environment (e.g., making a cup of tea in a hotel room), we have to search visually for the objects needed to perform the task (Ballard et al., 1992; Epelboim et al., 1995; Foerster et al., 2011; Foerster and Schneider, 2015b). In contrast, when acting in a familiar context, LTM can directly control gaze shifts to consecutive target objects in sequence, especially if the performed task is well-practiced. (Epelboim et al., 1995; Foerster et al., 2011, 2012; Foerster and Schneider, 2015b). As each of these task-driven gaze shifts is obligatorily preceded by a covert shift of attention (Deubel and Schneider, 1996), LTM controls for a sequence of attention and gaze shifts in this case. LTM-based attention and gaze control can be acquired through practice because sensorimotor routine tasks typically consist of fixed task elements that are repeated in a constant environment (e.g., making a cup of tea in your home kitchen). In this case, the sequence of perceptual input as well as of motor actions can be learned and automatized (see Robertson, 2007; Schuck et al., 2012; Schwarb and Schumacher, 2012 for perceptual vs. motor aspects of sequence learning and for the question whether sequences are learned on an item-to-item basis). However, sometimes sensorimotor routines have to be adapted to changing task elements or environments. In this case, the LTM-based mode of covert and overt (saccade) attentional selection has to be modified.

How is attentional selection modified if LTM-based expectations about probable object locations are no longer valid? If target objects are no longer at expected locations, visual search has to be reinitiated. Interestingly, if only a few target objects within a manual action sequence are unexpectedly displaced, visual search is performed even while having to act on unchanged targets in the sequence (Foerster and Schneider, 2015b). In Foerster and Schneider (2015b), participants had to click on eight numbered shapes in ascending sequence on a computer screen while eye movements were recorded. After having worked on a constant target position arrangement for 60-prechange trials, numbers 3 and 6 switched position. This action-sequence affecting change caused searching fixations while acting on the new located numbers, but also while acting on the consecutive non-displaced number 4. Eye-cursor coordination was even disturbed while acting on nearly any later target. These results imply that it is not possible to switch instantaneously back to the LTM-based mode of attention once it has been disturbed, even if this would be efficient for motor control. Instead, spatial changes that influence sub-actions of a sensorimotor action sequence cause a regression from an LTM-based mode of attentional selection to visual search beyond the change-affected sub-actions. In line with this result, further studies have shown that humans prefer visual information over memory information for action control in case of little automatization or a requirement for flexible behavior (Droll and Hayhoe, 2007; Patsenko and Altmann, 2010). However, while we have to adapt the mode of selection and manual action to target location changes in the environment, unexpected but action-irrelevant changes in target appearance do not necessarily afford a modification in selection and behavior. Nevertheless, processing such violations to LTM-based expectations about the task material might nevertheless have effects on covert and overt spatial attention allocation as well as manual action control, e.g., due to surprise (Horstmann and Herwig, 2015; Horstmann et al., 2016).

In Foerster and Schneider (2015b), expectation-discrepant shape changes of action targets (switch of shapes surrounding numbers 3 and 6) did neither affect eye movements, nor cursor performance arguing that LTM-based attentional selection was not disturbed by the action-irrelevant change. However, other studies have shown that non-spatial expectation-discrepant feature changes capture attention (Schützwohl, 1998; Horstmann, 2002, 2005). When a distractor has an unexpected feature, responding to the target slows down (Schützwohl, 1998) arguing that the expectation-discrepant distractor captures attention. Even if the target instead of a distractor appears with an expectation-discrepant feature, response slowing is often found (Horstmann, 2002, 2005, 2015). It has been argued that attention is allocated to the task-irrelevant surprising feature of the target instead to the feature that has to be reported (Horstmann, 2015). In line with this idea, gaze latency to a target with an unexpected color is shorter than to a target with an expected color, and fixations dwell longer on the first than on the latter (Horstmann and Herwig, 2015). An expectation-discrepant non-spatial feature seems to capture the eyes fast and binds attention thereafter – oculomotor capture. In real-world scenes, scene-inconsistent or otherwise expectation-discrepant objects are not only longer fixated, but also more frequently revisited (Loftus and Mackworth, 1978; Hollingworth and Henderson, 2002; Võ and Henderson, 2009; Võ et al., 2010) – a kind of second-order oculomotor capture. It seems that the surprising feature is rechecked repeatedly after having noticed it for the first time – a check-after-surprise mode of attentional selection.

Why is a check-after-surprise mode of attentional selection frequently applied during visual search (Loftus and Mackworth, 1978; Võ and Henderson, 2009; Võ et al., 2010; Horstmann, 2015; Horstmann and Herwig, 2015), but has not been found during sensorimotor control (Foerster and Schneider, 2015b)? In visual search, attention allocation is sensory-based, i.e., all visual objects and their features are potentially important to solve the task because the target can be anywhere in the visual environment. When performing a specific well-practiced sequential sensorimotor task, however, target features and locations are typically constant, so that LTM determines where-to-attend and where-to-look in sequence. Task-irrelevant objects and features are usually very effectively ignored (Land et al., 1999; Land and Hayhoe, 2001; Hayhoe et al., 2003; Droll et al., 2005; Foerster et al., 2011; Belardinelli et al., 2015). Thus, an expectation-discrepant but task-irrelevant feature seems to be effectively ignored in such tasks.

However, there are reasons to believe that a check-after-surprise mode of attentional selection could be useful during sequential sensorimotor control. In such tasks, changes of any kind might signal an unpredictable environment. Moreover,

features without relevance in a specific task might become relevant for another related task. When walking a well-known route for shopping, it would be beneficial if attention would be captured by a road closure taking part behind, so that the way back can be planned efficiently. In summary, there is experimental evidence and arguments that speak against as well as in favor of adopting a check-after-surprise mode of attentional selection after task-irrelevant changes during sequential sensorimotor control.

A criterion that might determine whether task-irrelevant changes are noticed and modify attention is their relationship to the task-relevant objects of the task. In Foerster and Schneider (2015b), the target shapes were neither action-defining, nor in any other respect relevant throughout the experiment. Although, the eight individual shapes were obligatorily connected to the eight action-defining target numbers, the sensorimotor sequence could have been learned and executed equally well without the redundant shape information. Therefore, the shapes in the number-clicking task had no informational value for sensorimotor task control and could be completely ignored from the very first trial on. Correspondingly, the shape changes did not capture attention. However, a task-irrelevant change should be processed if it is related to an action-defining feature such as the appearance of a sign instructing your behavior (e.g., different looking traffic signs in a foreign country). Such task-irrelevant expectation violations might therefore initiate a check-after-surprise mode of attentional selection also during sensorimotor control.

Here, it was investigated whether and how a check-after-surprise mode of attentional selection is applied in a well-practiced manual sequence after a task-irrelevant change that is bound to an action-defining feature. In a computerized version of the number-connection test, participants had to click as fast as possible with a mouse cursor in ascending sequence on nine spatially distributed numbered circles on a computer screen. Eye movements were recorded as proxy for attentional selection based on the fact that a covert shift of attention obligatorily precedes every saccade (Deubel and Schneider, 1996). To ensure that an LTM-based mode of attentional selection was used prior to the introduced change, participants had to work on a constant configuration of numbered circles throughout 65 prechange-trials. In 20 successive change-trials, the task-irrelevant font of number 4 was changed. In 15 final reversion-trials, the originally presented and learned font was used again. The hypothesis is that the font change on the number is processed because the identity of the number is action-defining, as it specifies the position of the action target in the sequence – and had to be attended to learn the sensorimotor sequence. Thus, participants should be surprised and check for the new appearance of the number 4 after having noticed the font change. The aim of the study was to reveal at which moment within the sensorimotor sequence attention is captured by the expectation-discrepant number font and for how long it is revisited within the sensorimotor sequence as well as across several trials. Is the change noticed when having to act on the changed number or already when acting on prior targets in the sequence? Will the changed target 4 be checked only while having to act on it or also after having clicked on it successfully? Will the new appearance of the target elicit checking fixations even in subsequent trials? How fast can the reversion to the originally learned display with the originally learned fonts for all number be processed? Are eye movements and manual actions affected differentially by the change? These questions are important to understand how covert attention, gaze, and manual sequences are planned, preprogrammed, executed, and updated during sensorimotor control of well-practiced sequential manual actions.

MATERIALS AND METHODS

Participants

Twenty right-handed students (8 males and 12 females, average age 25 years) from Bielefeld University, Germany, participated in the study after having provided written informed consent. All participants reported normal visual acuity, were naïve with respect to the purpose of the study, and were paid for their participation. The study was approved by the Committee for Ethics at Bielefeld University (EUB) and performed in accordance with the approved guidelines.

Apparatus and Stimuli

The experiment took place in a dimply lit room and stimuli were displayed on a 19-inch color CRT monitor (ViewSonic Graphics Series G90fB using an ATI Radeon HD 2400 Pro graphics card) with a refresh rate of 100 Hz and a spatial resolution of 1024 pixels × 768 pixels extending 36 cm × 27 cm. Viewing distance was fixed with a chin-and-forehead rest at 71 cm. The experiment was controlled by the Experiment Builder software (SR Research, Ottawa, ON, Canada) on a Dell Optiplex 755 computer. The right gaze position was recorded with 1000 Hz by an EyeLink 1000 tower-mounted eye tracker (SR Research, Ottawa, ON, Canada). The computer mouse and keyboard were used as well as an extra-large mouse pad (32 cm × 88 cm). Color and luminance were measured in CIE Lxy coordinates using an X-Rite i1 Pro spectrophotometer.

All stimuli were displayed on a gray background (RGB 204, 204, 204; $L = 78.9$ cd/m^2, $x = 0.29$, $y = 0.30$). The mouse cursor was a black dot of 0.43 degrees of visual angle (°v.a.) in diameter (RGB 0, 0, 0; $L = 0.3$ cd/m^2, $x = 0.32$, $y = 0.33$). The target stimuli consisted of nine black numbered circles (circle diameter of 2.04°v.a.; bold type Arial numbers of font size 35 which equals to app. 0.96°v.a. height and 0.62°v.a. width, number 4 also in bold type MV Boli in some trials, color and luminance identical to the cursor). Circle number 1 was centered on the computer screen. The spatial layout of the remaining eight numbered circles was designed by randomly choosing locations within the outer fields of an imagined 3 × 3 grid with the prerequisite that the circles had a minimal distance of 2.04° v.a. to each other (border-to-border) as well as to the screen border. The spatial layout of the nine numbered circles was constant throughout the experiment.

Procedure

Participants first read the instruction on the computer screen. They were asked to click on nine numbered circles in ascending

order as fast as possible. A nine-point eye-tracking calibration and validation procedure followed. Only calibrations with an averaged accuracy below 1.0°v.a. were accepted. The first trial was announced as an example trial and thus not included in the analyses. The experiment consisted of a 65-trials prechange-acquisition phase (example trial excluded), a 20-trials change phase, and a 15-trials reversion phase. While the font of number 4 was Arial throughout prechange and reversion phase, it appeared in the font MV Boli throughout the change trials (**Figure 1**). All other numbers were displayed in Arial font throughout the experiment. A click was counted as correct within a diameter of 3.06°v.a. around a target's center. A correct click was followed by a high-pitched tone. After all nine targets had been clicked in the correct sequence, a feedback display signaled the trial-completion time. A calibration check preceded each trial via a central fixation on a black ring (0.45°v.a. outer size, 0.11 °v.a. inner size). Calibration was repeated if necessary. After every block of 11 trials, a display informed participants about the number of completed and total experimental blocks. Participants started each block and trial by pressing the space bar. After having finished the last experimental trial, participants were asked whether they had noticed something peculiar. They used the keyboard to type in their answer. Subsequently, they were informed that indeed something was peculiar in the experiment and were asked to indicate which of 10 numbered statements did apply to the experiment (statements can be seen in **Table 1**). Selection was performed by typing in the selected statement numbers. All participants completed the experiment within 40 min. The participant with the fastest best time earned 2€ extra.

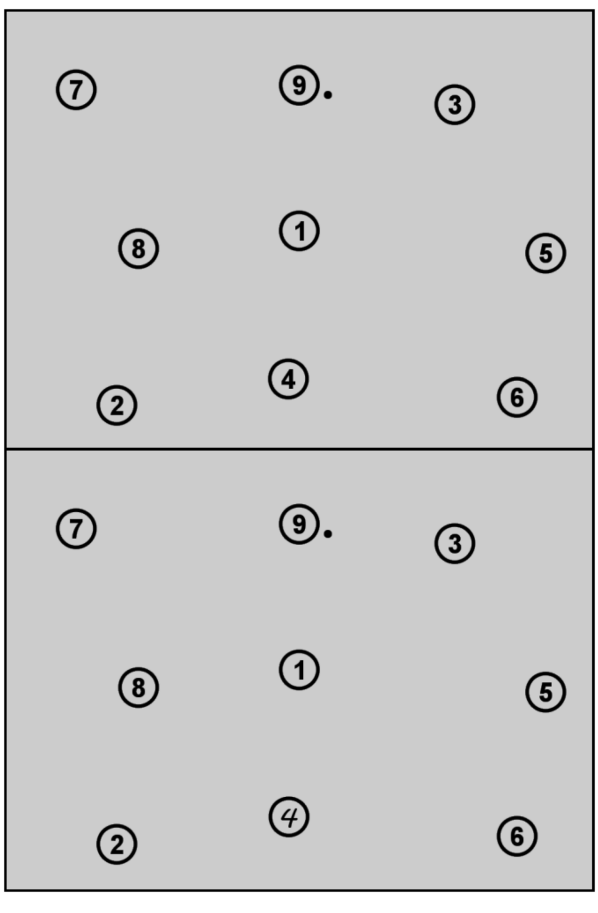


FIGURE 1 | Display during the clicking task in the prechange (top), the change (bottom), and the reversion (top) phase. The black dot near number nine displays the mouse cursor.

Analysis

The following dependent variables were analyzed: Trial-completion times, number of errors and fixations, scanpath and cursor-path lengths, as well as eye–cursor distance. The SR Research EyeLink Data Viewer software's implemented default velocity algorithm was used to detect fixations (not a blink, <30°v.a./s velocity and <8,000°v.a./s^2 acceleration). Scanpath and cursor-path lengths were calculated as 100-Hz cumulative inter-sample distances. Eye–cursor distance was calculated as 100-Hz intra-sample distance.

To reveal whether LTM-based attentional control was built up over the course of the prechange phase, analyses of variance (ANOVAs) studied the state of learning through the first five prechange blocks (1–11, 12–22, 23–33, 34–44, and 45–55). To analyze the effects of the font change, paired *t*-tests were conducted to compare the very first change trial (trial 66) to the prechange baseline consisting of the average of the last ten prechange trials (56–65). For fine-grained analyses, further within-subject variables were sub-action (1–9), location (1–9), and fixation type (searching, guiding, and checking). Fixation types were defined according to their landing positions (Epelboim et al., 1995; Land and Tatler, 2009; Foerster and Schneider, 2015a,b): fixations on any upcoming target (except the current target) as searching, fixations on a current target as guiding, and fixations on any completed target as checking (interest area of 3.06°v.a. diameter). To analyze how long the effects of the changed font might last when repeating the changed display, change trials 2–5 were also compared to prechange baseline with paired *t*-tests. To reveal whether the reversion to the originally learned display had any effects on performance and gaze control, paired *t*-tests were used

TABLE 1 | The English translation of the 10 numbered statements participants could indicate as applicable to the experiment as well as the number of choices (right column).

(0) The size of one or several stimuli was changed.	0
(1) One had to click spatially displaced on one or several stimuli.	1
(2) The order was changed.	3
(3) The font of one or several numbers was changed.	18
(4) One or several stimuli were spatially displaced.	3
(5) The shape of one or several stimuli was changed.	0
(6) The size of one or several numbers was changed.	2
(7) Clicks were not always accepted.	2
(8) One or several numbers was spatially displaced.	2
(9) One or several numbers were missing.	0

to compare the very first reversion trial to the prechange baseline. Violations of sphericity were corrected by using the Greenhouse-Geisser ε (uncorrected degrees of freedom are provided to facilitate reading). A chance level of 0.05 was applied.

RESULTS

This section is divided into four parts. First, it is analyzed whether participants adopted an LTM-based attention mode over the course of the prechange phase (five blocks). Second, I report the effects of the unexpected font change on manual performance and eye movements to reveal whether there was a shift in the applied mode of attentional selection, i.e., from an LTM-based mode to a check-after-surprise mode of attentional selection. Third, I report the effects on attentional control by several repetitions of the changed display as well as by the reversion to the originally learned display. The third investigation will reveal how long the surprising font change affected manual and eye movement parameters before an LTM-based mode of attentional control was reinitiated as well as whether the reversion to the prior font was as surprising in terms of modifications of gaze and manual action parameters as the initial font change. Finally, the answers to the explicit awareness questions will be summarized.

Prechange Phase: Acquisition of a LTM-Based Mode of Attentional Selection

Did participants adopt an LTM-based mode of attentional selection within the prechange phase? Over the course of the first five prechange blocks, trial completion time, number of fixations, cursor-path and scanpath length, and eye-cursor distance decreased as is typical for sensorimotor learning [**Figures 2A,B**; time: $F(4,76) = 48.38$, $p < 0.001$; linear trend $F(1,19) = 75.63$, $p < 0.001$; fixations: $F(4,76) = 41.99$, $\varepsilon = 0.56$, $p < 0.001$; linear trend $F(1,19) = 71.42$, $p < 0.001$; cursor-path: $F(4,76) = 23.28$, $\varepsilon = 0.53$, $p < 0.001$; linear trend $F(1,19) = 33.07$, $p < 0.001$; scanpath: $F(4,76) = 30.00$, $\varepsilon = 0.48$, $p < 0.001$; linear trend $F(1,19) = 48.60$, $p < 0.001$; eye-cursor distance: $F(4,76) = 6.60$, $\varepsilon = 0.53$, $p < 0.01$; linear trend $F(1,19) = 13.11$, $p < 0.01$]. An ANOVA on the number of fixations with block and fixation type as within-subject variables revealed significant main effects of block and type as well as a significant interaction [block: $F(4,76) = 29.15$, $\varepsilon = 0.54$, $p < 0.001$; type: $F(2,38) = 158.39$, $\varepsilon = 0.62$, $p < 0.001$; block by type: $F(8,152) = 4.92$, $\varepsilon = 0.27$, $p < 0.05$]. All types of fixations decreased significantly in the course of the prechange phase [**Figure 2C**; searching fixations: $F(4,76) = 4.10$, $\varepsilon = 0.64$, $p < 0.05$, linear trend $F(1,19) = 7.31$, $p < 0.05$; guiding fixations: $F(4,76) = 3.75$, $\varepsilon = 0.46$, $p < 0.05$, linear trend $F(1,19) = 4.90$, $p < 0.05$; checking fixations: $F(4,76) = 33.40$, $\varepsilon = 0.41$, $p < 0.001$, linear trend $F(1,19) = 45.14$, $p < 0.001$]. On average, significantly more guiding fixations were performed than searching and checking fixations, and more checking than searching fixations (all ps < 0.001). During the fifth prechange block, participants performed on average 8.95 guiding, 0.87 checking, and 0.27 searching fixations per trial. Guiding the hand (here cursor) sequentially with approximately one fixation to each target on an effective path is a typical characteristic of LTM-based attentional selection for sensorimotor control (Foerster et al., 2011, 2012; Foerster and Schneider, 2015a,b). None of the dependent variables was significantly different across blocks 4 and 5. Thus, a first plateau of gaze and manual action performance seemed to be reached after the 4th block.

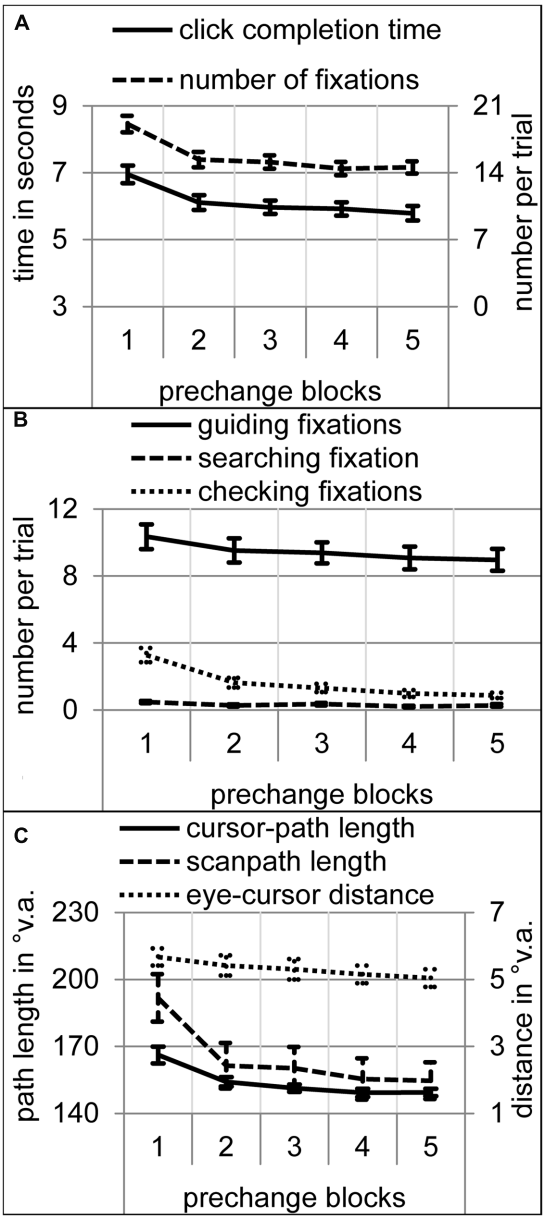


FIGURE 2 | Performance and eye movement measures over the course of the five prechange blocks. Error bars represent standard error of the means. **(A)** Click completion time in seconds and number of fixations per trial. **(B)** Number of searching, guiding, and checking fixations per trial. **(C)** Cursor-path and scanpath length as well as eye-cursor distance in °v.a.

First Task-Irrelevant Change Trial: Shift to a Checking-after-Surprise Mode of Attentional Selection

How did participants allocate their overt attention within the sensorimotor sequence, when number 4 appeared unexpectedly in another font? To answer this question, the dependent variables of the prechange baseline (last ten prechange trials) were compared to the very first change trial (**Figure 3**). The completion time of the change trial was significantly longer than in the prechange baseline [$t(19) = 2.20$, $p < 0.05$, **Figure 3A**]. In addition, participants performed more fixations [$t(19) = 4.72$, $p < 0.001$, **Figure 3C**] during the change trial. Number of errors and cursor-path length was not significantly affected by the font change [$t(19) = 0.17$, $p = 0.87$, **Figure 3B** and $t(19) = 1.00$, $p = 0.33$, **Figure 3D**, respectively]. However, scanpaths length and eye-cursor distance was larger when acting on the changed than the learned prechange display [$t(19) = 3.43$, $p < 0.01$, **Figure 3E**, and $t(19) = 3.31$, $p < 0.01$, **Figure 3F**, respectively].

An analysis of the number of the different fixation types revealed that significantly more searching and checking fixations were performed during the change trial [$t(19) = 2.68$, $p < 0.05$ and $t(19) = 4.28$, $p < 0.001$, respectively], but not significantly more guiding fixations [$t(19) = 1.41$, $p = 0.17$]. For the number of checking fixations, the interaction between condition (prechange vs. change) and location (1–9) was significant [$F(8,152) = 22.48$, $\varepsilon = 0.18$, $p < 0.001$]. The additional checking fixations were exclusively directed to number 4 [$t(19) = 5.24$, $p < 0.001$, **Figure 4A**]. Obviously, the changed font of number 4 caused attentional and oculomotor revisiting. Also for the number of guiding fixations, the condition by location interaction reached significance [$F(8,152) = 2.11$, $\varepsilon = 0.83$, $p < 0.05$]. More guiding fixations were performed on the changed number 4 [$t(19) = 3.26$, $p < 0.01$, **Figure 4B**]. The increase in searching fixations was not accompanied by a significant condition-by-location interaction [$F(8,152) = 1.15$, $\varepsilon = 0.43$, $p = 0.34$]. Thus, the increase in searching fixations was not concerned with a specific location (**Figure 4C**).

During which sub-action of the sensorimotor sequence was participants' attention captured by the expectation-discrepant appearance of number 4? To answer this question, analyses of variance with click action (1–9) and condition (prechange vs. change) as within-subject factors were calculated for the

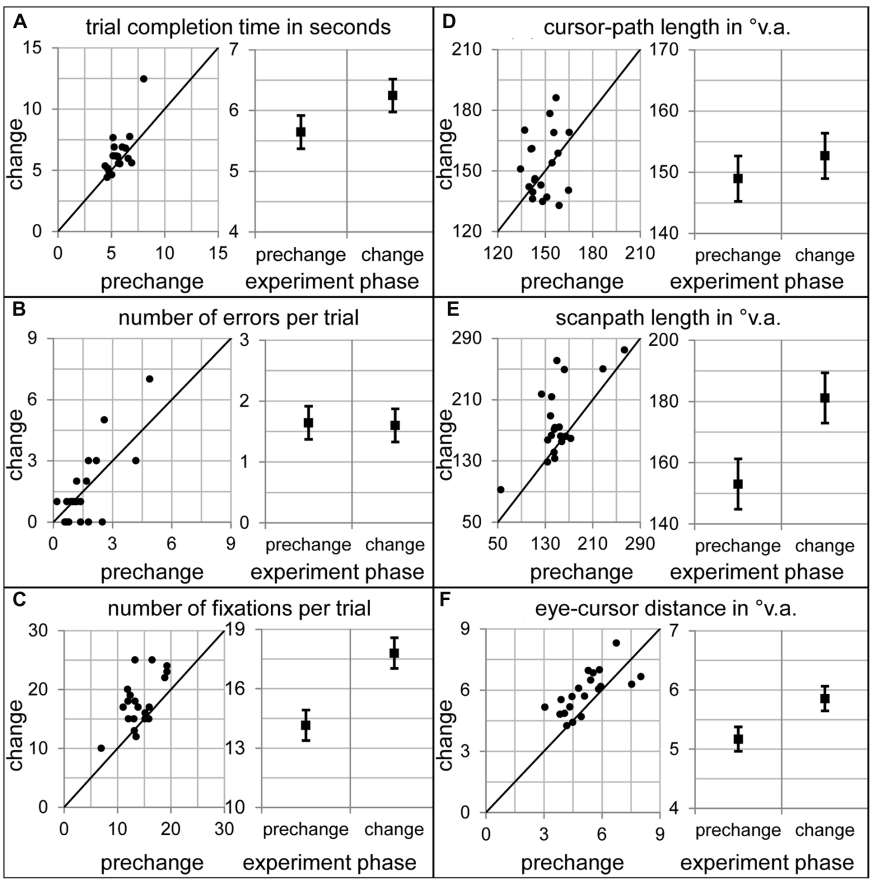


FIGURE 3 | Dependent variables during prechange baseline and during the first change trial. The left diagrams of all panels show the 20 participants' individual data. The right diagrams of all panels show the sample means and the standard error of the paired mean differences. **(A)** Click completion time in seconds. **(B)** Number of errors per trial. **(C)** Number of fixations per trial. **(D)** Cursor-path length in °v.a. **(E)** Scanpath length in °v.a. **(F)** Eye-cursor distance in °v.a.

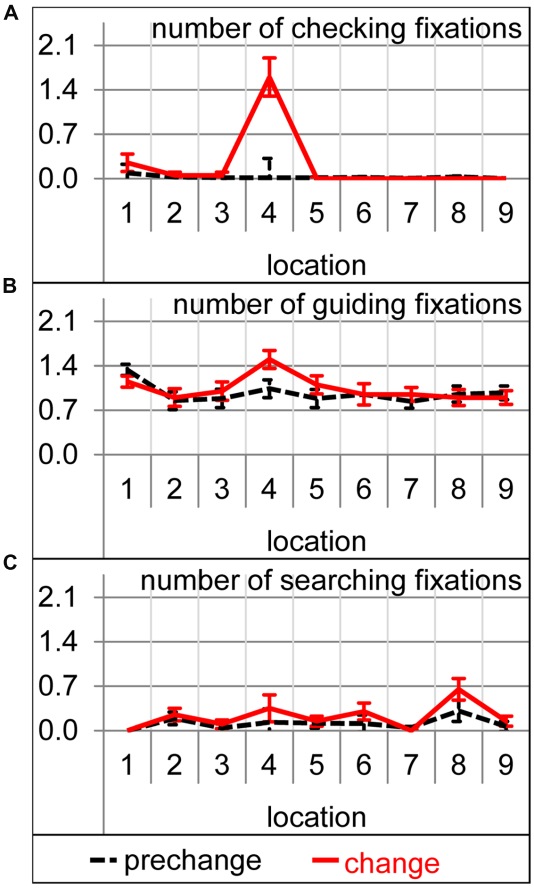


FIGURE 4 | Number of the checking (A), guiding **(B)**, and searching **(C)** fixation per trial on each of the nine target locations during prechange baseline (broken black lines) and the first change trial (solid red lines). Error bars represent standard errors of the paired difference between prechange and change for each location.

number of each fixation type. For the number of checking fixations, the interaction between condition and click action was significant [$F(8,152) = 5.98$, $\varepsilon = 0.39$, $p < 0.01$] as were both main effects [condition: $F(1,19) = 18.31$, $p < 0.001$; action: $F(8,152) = 6.29$, $\varepsilon = 0.40$, $p < 0.01$]. Significantly more checking fixations were performed during click actions 6 [$t(19) = 2.83$, $p < 0.05$], 8 [$t(19) = 2.93$, $p < 0.01$], and 9 [$t(19) = 3.05$, $p < 0.01$], and marginally during click action 5 [$t(19) = 2.03$, $p = 0.06$; **Figure 5B**]. The analysis of guiding fixations per click action is identical to the analysis of guiding fixation per location, as guiding fixations are always concerned with the current action target location. As already mentioned above, the changed font caused more guiding fixations for number 4 and thus also during click action 4 (**Figures 4B** and **5C**). The analysis of the number of searching fixations resulted in no significant interaction between condition and click action [$F(8,152) = 1.24$, $\varepsilon = 0.46$, $p = 0.30$], but a significant main effects of condition [$F(1,19) = 7.20$, $p < 0.05$] and action [$F(8,152) = 3.72$, $\varepsilon = 0.47$, $p < 0.01$). Thus, some sub-actions afforded generally more searching than others. However, the increase in the searching behavior was not concerned with a specific sub-action of the sensorimotor sequence. Also for scanpaths length, the interaction of condition and click action did not reach significance [$F(8,152) = 1.24$, $\varepsilon = 0.46$, $p = 0.30$], but the main effects of condition [$F(1,19) = 11.76$, $p < 0.01$, **Figure 5E**] and action [$F(8,152) = 11.45$, $\varepsilon = 0.50$, $p < 0.001$] did. Scanpaths were generally longer during some click actions, and prolonged due to the change. However, their prolongation was not concerned with a specific click action. While eye movement parameters were strongly affected by the font change, manual performance did not suffer remarkably. Trial completion time was prolonged (see above). The increase in completion time was not concerned with a specific click action [action by condition interaction: $F(8,152) = 1.63$, $\varepsilon = 0.38$, $p = 0.19$, **Figure 5A**]. Cursor-path length did not at all increase as already mentioned above (**Figure 3D**, also **Figure 5D** per click action). This dissociation between eye and cursor movements was confirmed by the increased distance between eye and cursor. Not only the main effects of condition [$F(1,19) = 10.96$, $p < 0.01$] and action [$F(8,152) = 18.86$, $\varepsilon = 0.52$, $p < 0.001$], but also the interaction was significant [$F(8,152) = 3.35$, $\varepsilon = 0.53$, $p < 0.05$]. Eye-cursor distance was increased across click actions 6–9 [6: $t(19) = 2.81$, $p < 0.05$; 7: $t(19) = 2.19$, $p < 0.05$; 8: $t(19) = 2.64$, $p < 0.05$; 9: $t(19) = 3.58$, $p < 0.01$, **Figure 5F**]. Thus, eye-cursor coupling was disturbed after having acted on the font-changed number 4, but not before.

In summary, when the task-irrelevant font of the action-defining number 4 changed unexpectedly in the well-practiced sensorimotor task, the LTM-based mode of attentional selection was replaced by a check-after-surprise mode of attentional selection. Moreover, while the surprising feature frequently attracted the eyes after it had been acted on, the hand continued the task without much interference. This result revels that eye-hand coupling is loosened in the service of maintaining manual performance.

Change Repetition and Reversal: Reinitiation of the LTM-Based Mode of Attentional Selection

The effects on attentional control by several repetitions of the changed display as well as by the reversion to the originally learned display can be found here. Investigating the effects elicited by a repetition of the font-changed display can reveal how long the unexpected font was rechecked with the eyes during task performance until an LTM-based mode of attentional control was reinitiated. Planned *t*-tests were performed to compare the four subsequent change trials (trials 67–70) individually to the prechange baseline. The first repetition of the changed display resulted in a longer trial completion time [$t(19) = 2.31$, $p < 0.05$], more fixations [$t(19) = 5.20$, $p < 0.001$], which were searching fixations [$t(19) = 4.18$, $p < 0.001$], longer scanpaths [$t(19) = 3.31$, $p < 0.01$], and a larger eye-cursor distance [$t(19) = 2.75$, $p < 0.05$]. More fixations were also performed during the second repetition of the changed display [$t(19) = 2.54$, $p < 0.05$], but this time more guiding fixations [$t(19) = 3.00$, $p < 0.01$]. All other dependent variables did no longer differ significantly to

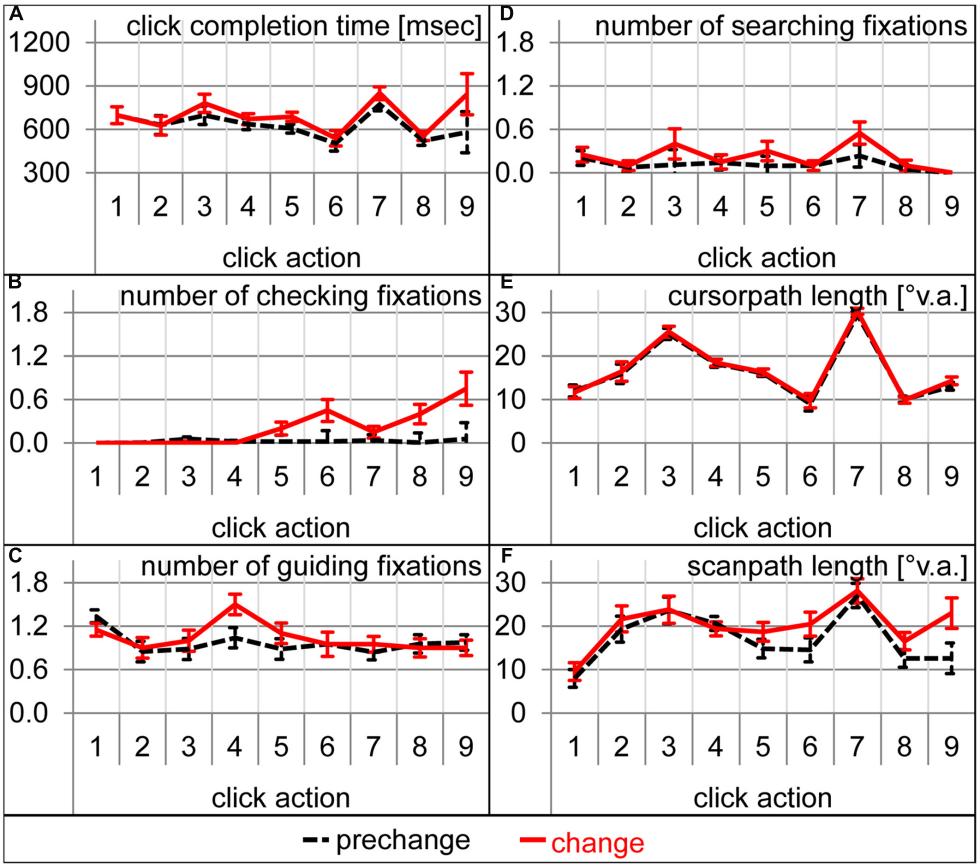


FIGURE 5 | Dependent variables during each of the nine clicking actions during prechange baseline (broken black lines) and the first change trial (solid red lines). Error bars represent standard errors of the paired difference between prechange and change for each click action. **(A)** Click completion time in milliseconds. **(B)** Number of checking fixations. **(C)** Number of guiding fixations. **(D)** Number of searching fixations. **(E)** Cursor-path length in °v.a. **(F)** Scanpath length in °v.a.

prechange baseline. Thus, the surprising deviant font did affect performance and gaze control only up to two repetitions before participants worked in a LTM-based mode of attentional control again.

Did the reversion to the previously learned Arial-font display elicit the same surprise effect as the initial font change? The very first reversion trial (86) differed significantly from the prechange baseline only in trial completion time [$t(19) = 2.60$, $p < 0.05$]. However, participants were not slower, but faster during the reversion trial, perhaps due to further motor refinement over the course of the change phase. Thus, the reversion to the originally learned display did not elicit any check-after-surprise effects in terms of gaze performance changes.

Explicit Awareness of the Font Change

In order to reveal, whether participants were explicitly aware of the font change, they were asked after the experiment whether they had noticed something peculiar. Ten of the 20 participants spontaneously reported that the font of number 4 did change within the experiment. When participants had to select a noticed change from ten presented alternatives (see **Table 1**), 18 of the 20 participants selected the font change. Nine participants indicated further statements to be true. Most participants seemed to have noticed the font change explicitly. The occasional entry of further observed changes might have been encouraged by the permission of multiple selections and a natural suspicion of psychology students with respect to experimental manipulations.

DISCUSSION

In the present study, it was investigated whether and how attentional selection of action targets for sequential motor routines is modified when confronted with task-irrelevant expectation-violations. Although environmental changes that are not relevant for the current task do not require action modification, they are nevertheless unexpected and might therefore influence overt attentional selection and manual control. Especially if task-irrelevant aspects of action-defining features are changed, attentional selection based on LTM expectations might be disturbed. The hypothesis was that a task-irrelevant change on an action-defining feature of a target should

lead to rechecking of the expectation-violating object. Revisits of the changed target could be purely oculomotor or also manual, which has different consequences for eye-hand coupling. It was investigated how long possible effects of the change would last within the action sequence as well as when repeatedly displaying the changed target.

In a computerized version of the number-connection test, participants clicked as fast as possible with a mouse cursor in ascending sequence on nine spatially distributed numbered circles on a computer screen, while gaze was recorded. Participants had to work on a constant configuration of numbered circles throughout 65 prechange-trials. In 20 successive change-trials, the font of number 4 changed. In 15 final reversion-trials, the originally learned font was used again. Results revealed that the font-changed number 4 captured attention and eye movements as soon as number 4 had to be acted on. Cursor movements, however, were not at all affected. The asymmetry between eye and cursor effects was reflected by an enlarged eye-cursor distance throughout the remaining trial. The effects lasted for up to two repetitions of the changed target display, but did not reoccur when reverting to the originally learned display.

In the following, the results are discussed with respect to the involvement of a sensory-based vs. memory-based control mode of attention and eye movements when performing well-practiced sequential sensorimotor actions. Afterward, the checking-after-surprise gaze effect will be dissociated from gaze effects caused by the need for a modification of a learned sensorimotor sequence. Finally, the limits of eye-hand coupling are discussed.

Sensory-Based versus Memory-Based Control of Attention and Gaze

When having to determine where to attend next to achieve an ongoing task, different sources of information can be used. Sensory information is weighted according to its task-relevance (Bundesen et al., 2011; Bundesen and Habekost, 2014; Poth et al., 2014). Attention and gaze can then be shifted to the location containing the most relevant information (e.g., highest attentional weight) for the current task (Wischnewski et al., 2010; Schneider, 2013). Alternatively, task-related memory can be used to shift attention and gaze directly to a retrieved target position without the need to process visual features (Foerster et al., 2011, 2012, Jiang et al., 2013, 2014). Especially, when performing well-known sensorimotor actions, strong memory codes are used to direct attention and gaze in a task-dependent manner (Foerster et al., 2011, 2012). When switching on your bedside light in the dark, attention and gaze can be directly shifted to the light switch from LTM, allowing to perform the task even in complete darkness (cf. Foerster et al., 2012). Usually, both memory-based and sensory-based control of attention is applied to achieve a task.

Investigating which available sensory information is still modulating gaze control in a well-practiced sensorimotor task can reveal whether and how the contents of task sets are modulated throughout sensorimotor learning, such as relying more or even exclusively on LTM for action control and ignoring specific sensory information completely (Schneider and Shiffrin, 1977; Logan, 1988). In the present study, sensory-based attention and gaze control were still applied after extensive learning, even for constant sensory features that are no longer indicative for successful action control. The font change caused participants to frequently revisit the font-changed number 4, although the change did not afford a modification of the learned and memorized sensorimotor trajectory. Participants also noticed the font change explicitly. Conclusively, they did still process the sequence-defining identity of the numbers which is not separable from its font. Number identities were processed although LTM would have sufficed to determine and execute the clicking sequence. It seems that the task set was still defined according to the instruction to click the numbered circles in ascending sequence. Participants did not modulate the task set, e.g., into "click the learned sequence." Participants still used sensory information for gaze control and did not rely completely on learned spatial or motor codes (see Hikosaka et al., 1999; Rand et al., 2000; Richard et al., 2009 for the application of spatial and motor codes in sensorimotor sequences). Thus, even in well-practiced sensorimotor tasks, the available sensory information is still extracted and processed according to explicit task sets. Cursor movements, however, were not affected by the changed sensory information. This finding argues for a stronger contribution of spatial and motor codes for manual than for oculomotor control.

The continued application of sensory-based gaze control is in line with the observation that the eyes typically sample sensory information for a current sub-action just in time (Land et al., 1999; Hayhoe et al., 2003; Gajewski and Henderson, 2005). The currently important visual features and locations are extracted shortly before they are needed even if they could be recalled from memory (Gajewski and Henderson, 2005; Droll and Hayhoe, 2007; Foerster et al., 2011, 2012). Using the world as external memory saves memory load (O'Regan, 1992). Moreover, the current visual information is more reliable and richer in detail than an error-prone memory trace (Gray and Fu, 2004). Finally, revisiting action-relevant visual information ensures fast and efficient adaptation to environmental changes across repetitions if required (Adam et al., 2012).

Gaze Effects Caused by Surprise versus the Need for Sensorimotor Modification

In the present investigation, modification of the manual trajectory was not needed as the action-defining number identity was not changed, only its font. The font change detection did nevertheless affect gaze behavior. The changed gaze behavior in the present study does not reflect a consequence of the need for a sensorimotor modification in response to a trajectory change as was the case in Foerster and Schneider (2015b). Instead the changed oculomotor selection here represents a surprise reaction to the task-irrelevant expectation violation. There are three differences in the gaze effects caused by surprise vs. the need for sensorimotor modification.

First, when a detected change affords a modification of a well-practiced sensorimotor sequence, gaze control regresses from a LTM-based mode to a visual search mode (Foerster and

Schneider, 2015b). In such a case, not yet completed action targets are frequently visited in order to find the changed target locations or sequence. However, completed action targets are still nearly completely ignored, i.e., no checking fixations are performed. When an unexpected change elicits a surprise effect, oculomotor capture is observed, i.e., the surprising feature is fixated longer and is frequently revisited (Loftus and Mackworth, 1978; Hollingworth and Henderson, 2002; Võ and Henderson, 2009; Võ et al., 2010). Correspondingly, in the present study, the font-changed number 4 was frequently checked after having successfully clicked on it. Interestingly, this modulation of attentional selection was not accompanied by a change in the manually controlled cursor trajectory. This result pattern can also preclude that the effects are due to a lack of priming for the new font. Primed stimuli can be selected and responded to faster than unprimed stimuli (Meeter and Van der Stigchel, 2013; Kruijne and Meeter, 2016). However, the increase in clicking time in the present study was not limited to the font-changed number. In addition, the cessation of priming for the font-changed number would not predict a revisiting of this unprimed material, especially not after having clicked on it successfully.

Second, the effects on attentional selection caused by the need to modify a well-practiced sensorimotor sequence remains for several repetitions. Up to 15 trials were needed to fully integrate a two-numbers location switch in the 8-target sequence of Foerster and Schneider (2015b), so that gaze control showed again all characteristics of LTM-based selection. Surprise effects, however, are typically short-lived. Only the very first presentation of a deviant stimulus style resulted in longer reaction times in the study of Schützwohl (1998), while reaction times to repeated deviant styles were not significantly different from prechange baseline. Also the gaze effects elicited by the font change of the present study did result in relatively short-lasting effects (1–3 trials).

Third, the reinitiation of a previously learned sensorimotor sequence is accompanied by the same visual search gaze behavior as the first modification (Foerster and Schneider, 2015b). Contrastingly, the reversion to a familiar visual presentation elicits usually no response change indicative for a surprise in contrast to the change from a familiar to an unfamiliar stimulation (e.g., Schützwohl, 1998; Horstmann, 2002; Horstmann and Herwig, 2015). Surprising a person twice is difficult enough, especially with something that is already known. Correspondingly, reversion to the original font in the present study did neither impair performance nor affect gaze. It is likely, that even if a new deviant had been introduced, no further or a far smaller surprise effect would have arisen. The introduction of a deviant should change the expectation about possible task elements "once and for all" as Gaschler et al. (2015) put it. On the basis of these three considerations, the gaze modification observed in the present study constitutes a check-after-surprise effect.

The Limits of Eye-Hand Coupling

When performing a sensorimotor task, eye and hand movements are typically tightly coupled (Neggers and Bekkering, 2000; Beurze et al., 2009; Song and McPeek, 2009). Before the hand or a manipulated tool reaches a specific action target, gaze is shifted to the target position (Land et al., 1999; Hayhoe et al., 2003; Sailer et al., 2005; Foerster et al., 2011, 2012). Eye movements typically precede hand movements. First, visual information important for hand motor planning can be extracted, e.g., target size and its orientation (Prablanc et al., 1979, 1986, 2003; Prablanc and Martin, 1992; Paillard, 1996; Land et al., 1999; Crawford et al., 2004; Prado et al., 2005; Beurze et al., 2006). Second, even if the visual appearance of the target is known, motor performance should benefit from the well-learned eye to hand motor transformations (Gnadt et al., 1991; Henriques et al., 2003; Crawford et al., 2004; Flanagan et al., 2008). Third, gaze can possibly be used as deictic pointer for the eyes (Ballard et al., 1992, 1997; Neggers and Bekkering, 2001; Flanagan et al., 2008; Rosenbaum, 2010). Forth, fixating an action target might serve as retrieval cue for the required action on the target or upcoming subactions of the sensorimotor sequence (Laeng and Teodorescu, 2002; Johansson et al., 2011; Johansson and Johansson, 2013). However, eye and hand can be decoupled if explicitly required by the task. We can, for instance, simultaneously saccade to one location and reach to another. In this case, attention is allocated in parallel to the saccade and reaching target prior to motor initiation (Jonikaitis and Deubel, 2011). An unanswered research question is how eye and hand movements are selected during sensorimotor actions in which the eyes are not arbitrarily restricted but assist the manual actions as is typical in real-world situations. It is possible that a common mechanism selects eye and hand target positions in this case (Schneider, 1995; Deubel and Schneider, 2003). Alternatively, eye and hand target positions might nevertheless be selected by different attentional mechanisms. If the latter is the case, spontaneous decoupling of eye and hand movements could occur even if the task does not afford a decoupling. In the present study, such a decoupling of eye and hand movements was observed. While the eyes revisited the font-changed number frequently, the hand-controlled cursor proceeded to move sequentially and with a similar speed to the remaining action targets. However, each target click was still preceded by a target fixation. Thus, eye-hand coupling was only partly abandoned for the sake of performance maintenance. While, not every saccade target selection was coupled to a cursor target selection, every manual target selection (click) was preceded by a saccade on the selected target. This is a nice analogy to the well-known finding that attention can be shifted covertly without moving the eyes, while each saccade is obligatorily preceded by a covert shift of attention (Deubel and Schneider, 1996). Future studies have to identify whether the results might generalize to tasks with other requirements, e.g., higher accuracy requirements when acting on objects with varying shapes or real-world interactions with three-dimensional objects.

Summary

In the present study, a task-irrelevant target feature change in a well-practiced sensorimotor task affected gaze behavior, while manual performance was hardly changed. Although target features can be retrieved from LTM in a well-practiced

sensorimotor task, target features that are action-defining according to the task set seem to be still visually processed. That is why the font change of the action-defining number target was detected in the present study. Detecting such action-defining feature changes allows flexible sensorimotor adaptation whenever needed. Although the detected font change did not require a sensorimotor adaptation in the present study, the deviant-font number was frequently revisited by the eyes. The violation of the learned LTM prediction elicited a surprise resulting in a checking mode of attention and gaze control for up to two repetitions of the deviant target display. Manual performance was hardly affected demonstrating that eye and hand movements can be efficiently decoupled in order to maintain a high level of task performance. Nevertheless, eye-hand coupling was still preserved for all target clicks displayed by target fixations guiding all successful mouse clicks. Therefore, an LTM-based mode of attention and gaze control is combined with a check-after-surprise mode after detecting task-irrelevant target feature changes while performing a well-practiced sensorimotor sequence.

AUTHOR CONTRIBUTIONS

RF designed and programmed the experiment, performed the data analysis, interpreted and discussed the results, and wrote the manuscript.

FUNDING

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). I acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.

ACKNOWLEDGMENT

I like to thank Werner X. Schneider who made crucial in-depth comments on the manuscript.

REFERENCES

Adam, J. J., Buetti, S., and Kerzel, D. (2012). Coordinated flexibility: how initial gaze position modulates eye-hand coordination and reaching. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 891–901. doi: 10.1037/a0027592

Ballard, D. H., Hayhoe, M. M., Li, F., Whitehead, S. D., Frisby, J. P., Taylor, J. G., et al. (1992). Hand-eye coordination during sequential tasks. *Philos. Trans. Biol. Sci.* 337, 331–339. doi: 10.1098/rstb.1992.0111

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behav. Brain Sci.* 20, 723–742. doi: 10.1017/S0140525X97001611

Belardinelli, A., Herbort, O., and Butz, M. V. (2015). Goal-oriented gaze strategies afforded by object interaction. *Vision Res.* 106, 47–57. doi: 10.1016/j.visres.2014.11.003

Beurze, S. M., de Lange, F. P., Toni, I., and Medendorp, W. P. (2009). Spatial and effector processing in the human parietofrontal network for reaches and saccades. *J. Neurophysiol.* 101, 3053–3062. doi: 10.1152/jn.91194.2008

Beurze, S. M., Van Pelt, S., and Medendorp, W. P. (2006). Behavioral reference frames for planning human reaching movements. *J. Neurophysiol.* 96, 352–362. doi: 10.1152/jn.01362.2005

Bundesen, C., and Habekost, T. (2014). "Theory of Visual Attention (TVA)," in *The Oxford Handbook of Attention*, eds S. Kastner and K. Nobre (Oxford: Oxford University Press), 1095–1121.

Bundesen, C., Habekost, T., and Kyllingsbæk, S. (2011). A neural theory of visual attention and short-term memory (NTVA). *Neuropsychologia* 49, 1446–1457. doi: 10.1016/j.neuropsychologia.2010.12.006

Crawford, J. D., Medendorp, W. P., and Marotta, J. J. (2004). Spatial transformations for eye-hand coordination. *J. Neurophysiol.* 92, 10–19. doi: 10.1152/jn.00117.2004

Deubel, H., and Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Res.* 36, 1827–1837. doi: 10.1016/0042-6989(95)00294-4

Deubel, H., and Schneider, W. X. (2003). Delayed saccades, but not delayed manual aiming movements, require visual attention shifts. *Ann. N. Y. Acad. Sci.* 1004, 289–296. doi: 10.1196/annals.1303.026

Droll, J. A., and Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 1352–1365. doi: 10.1037/0096-1523.33.6.1352

Droll, J. A., Hayhoe, M. M., Triesch, J., and Sullivan, B. T. (2005). Task demands control acquisition and storage of visual information. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1416–1438. doi: 10.1037/0096-1523.31.6.1416

Epelboim, J. L., Steinman, R. M., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C. J., et al. (1995). The function of visual search and memory in sequential looking tasks. *Vision Res.* 35, 3401–3422. doi: 10.1016/0042-6989(95)00080-X

Flanagan, J. R., Terao, Y., and Johansson, R. S. (2008). Gaze behavior when reaching to remembered targets. *J. Neurophysiol.* 100, 1533–1543. doi: 10.1152/jn.90518.2008

Foerster, R. M., Carbone, E., Koesling, H., and Schneider, W. X. (2011). Saccadic eye movements in a high-speed bimanual stacking task: changes of attentional control during learning and automatization. *J. Vis.* 11:9. doi: 10.1167/11.7.9

Foerster, R. M., Carbone, E., Koesling, H., and Schneider, W. X. (2012). Saccadic eye movements in the dark while performing an automatized sequential high-speed sensorimotor task. *J. Vis.* 12:8. doi: 10.1167/12.2.8

Foerster, R. M., and Schneider, W. X. (2015a). Anticipatory eye movements in sensorimotor actions: on the role of guiding fixations during learning. *Cogn. Process.* 16(Suppl. 1), 227–231. doi: 10.1007/s10339-015-0701-1

Foerster, R. M., and Schneider, W. X. (2015b). Expectation violations in sensorimotor sequences: shifting from LTM-based attentional selection to visual search. *Ann. N. Y. Acad. Sci.* 1339, 45–59. doi: 10.1111/nyas.12729

Gajewski, D. A., and Henderson, J. M. (2005). Minimal use of working memory in a scene comparison task. *Vis. cogn.* 12, 979–1002. doi: 10.1080/13506280444000616

Gaschler, R., Marewski, J. N., and Frensch, P. A. (2015). Once and for all - How people change strategy to ignore irrelevant information in visual tasks. *Q. J. Exp. Psychol.* 68, 543–567. doi: 10.1080/17470218.2014.961933

Gnadt, J. W., Bracewell, R. M., and Andersen, R. A. (1991). Sensorimotor transformation during eye movements to remembered visual targets. *Vision Res.* 31, 693–715. doi: 10.1016/0042-6989(91)90010-3

Gray, W. D., and Fu, W.-T. (2004). Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cogn. Sci.* 28, 359–382. doi: 10.1016/j.cogsci.2003.12.001

Hayhoe, M. M., Shrivastava, A., Mruczek, R., and Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *J. Vis.* 3, 49–63. doi: 10.1167/3.1.6

Henriques, D. Y. P., Medendorp, W. P., Gielen, C. C. A. M., and Crawford, J. D. (2003). Geometric computations underlying eye-hand coordination: orientations of the two eyes and the head. *Exp. Brain Res.* 152, 70–78. doi: 10.1007/s00221-003-1523-4

Hikosaka, O., Sakai, K., Lu, X., Nakahara, H., Rand, M. K., Nakamura, K., et al. (1999). Parallel neural networks for learning sequential procedures. *Trends Neurosci.* 22, 464–471. doi: 10.1016/S0166-2236(99)01439-3

Hollingworth, A., and Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 113–136. doi: 10.1037//0096-1523.28.1.113

Horstmann, G. (2002). Evidence for attentional capture by a surprising color singleton in visual search. *Psychol. Sci.* 13, 499–505. doi: 10.1111/1467-9280.00488

Horstmann, G. (2005). Attentional capture by an unannounced color singleton depends on expectation discrepancy. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1039–1060. doi: 10.1037/0096-1523.31.5.1039

Horstmann, G. (2015). The surprise-attention link: a review. *Ann. N. Y. Acad. Sci* 1339, 106–115. doi: 10.1111/nyas.12679

Horstmann, G., Becker, S., and Ernst, D. (2016). Perceptual salience captures the eyes on a surprise trial. *Attent. Percept. Psychophys.* 78, 1889–1900. doi: 10.3758/s13414-016-1102-y

Horstmann, G., and Herwig, A. (2015). Surprise attracts the eyes and binds the gaze. *Psychon. Bull. Rev.* 22, 743–749. doi: 10.3758/s13423-014-0723-1

Jiang, Y. V., Swallow, K. M., Rosenbaum, G. M., and Herzig, C. (2013). Rapid acquisition but slow extinction of an attentional bias in space. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 87–99. doi: 10.1037/a0027611

Jiang, Y. V., Won, B.-Y., and Swallow, K. M. (2014). First saccadic eye movement reveals persistent attentional guidance by implicit learning. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 1161–1173. doi: 10.1037/a0035961

Johansson, R., Holsanova, J., Dewhurst, R., and Holmqvist, K. (2011). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1289–1314. doi: 10.1037/a0026585

Johansson, R., and Johansson, M. (2013). Look here, eye movements play a functional role in memory retrieval. *Psychol. Sci* 25, 236–242. doi: 10.1177/0956797613498260

Jonikaitis, D., and Deubel, H. (2011). Independent allocation of attention to eye and hand targets in coordinated eye-hand movements. *Psychol. Sci.* 22, 339–347. doi: 10.1177/0956797610397666

Kruijne, W., and Meeter, M. (2016). Implicit short- and long-term memory direct our gaze in visual search. *Attent. Percept. Psychophys.* 78, 761–773. doi: 10.3758/s13414-015-1021-3

Laeng, B., and Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cogn. Sci.* 26, 207–231. doi: 10.1207/s15516709cog2602_3

Land, M. F., and Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Res.* 41, 3559–3565. doi: 10.1016/S0042-6989(01)00102-X

Land, M. F., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 1311–1328. doi: 10.1068/p2935

Land, M. F., and Tatler, B. W. (2009). *Looking and Acting*. New York, NY: Oxford University Press.

Loftus, G. R., and Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 4, 565–572. doi: 10.1037/0096-1523.4.4.565

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychol. Rev.* 95, 492–527. doi: 10.1037//0033-295X.95.4.492

Meeter, M., and Van der Stigchel, S. (2013). Visual priming through a boost of the target signal: evidence from saccadic landing positions. *Attent. Percept. Psychophys.* 75, 1336–1341. doi: 10.3758/s13414-013-0516-z

Neggers, S. F. W., and Bekkering, H. (2000). Ocular gaze is anchored to the target of an ongoing pointing movement. *J. Neurophysiol.* 83, 639–651.

Neggers, S. F. W., and Bekkering, H. (2001). Gaze anchoring to a pointing target is present during the entire pointing movement and is driven by a non-visual signal. *J. Neurophysiol.* 86, 961–970.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* 46, 461–488. doi: 10.1037/h0084327

Paillard, J. (1996). Fast and slow feedback loops for the visual correction of spatial errors in a pointing task: a reappraisal. *Can. J. Physiol. Pharmacol.* 74, 401–417. doi: 10.1139/y96-033

Patsenko, E. G., and Altmann, E. M. (2010). How planful is routine behavior? A selective-attention model of performance in the Tower of Hanoi. *J. Exp. Psychol. Gen.* 139, 95–116. doi: 10.1037/a0018268

Poth, C. H., Petersen, A., Bundesen, C., and Schneider, W. X. (2014). Effects of monitoring for visual events on distinct components of attention. *Front. Psychol.* 5:930. doi: 10.3389/fpsyg.2014.00930

Prablanc, C., Desmurget, M., and Gréa, H. (2003). Neural control of on-line guidance of hand reaching movements. *Progress Brain Res.* 142, 155–170. doi: 10.1016/S0079-6123(03)42012-8

Prablanc, C., Echallier, J. F., Komilis, E., and Jeannerod, M. (1979). Optimal response of eye and hand motor systems in pointing at a visual target. *Biol. Cybern.* 35, 113–124. doi: 10.1007/BF00337436

Prablanc, C., and Martin, O. (1992). Automatic control during hand reaching at undetected two-dimensional target displacements. *J. Neurophysiol.* 67, 455–469.

Prablanc, C., Pellison, D., and Goodale, M. A. (1986). Visual control of reaching movements without vision of the limb. *Exp. Brain Res.* 62, 293–302. doi: 10.1007/BF00238848

Prado, J., Clavagnier, S., Otzenberger, H., Scheiber, C., Kennedy, H., and Perenin, M. T. (2005). Two cortical systems for reaching in central and peripheral vision. *Neuron* 48, 849–858. doi: 10.1016/j.neuron.2005.10.010

Rand, M. K., Hikosaka, O., Miyachi, S., Lu, X., Nakamura, K., Kitaguchi, K., et al. (2000). Characteristics of sequential movements during early learning period in monkeys. *Exp. Brain Res.* 131, 293–304. doi: 10.1007/s002219900283

Richard, M. V., Clegg, B. A., and Seger, C. A. (2009). Implicit motor sequence learning is not represented purely in response locations. *Q. J. Exp. Psychol.* 62, 1516–1522. doi: 10.1080/17470210902732130

Robertson, E. M. (2007). The serial reaction time task: implicit motor skill learning? *J. Neurosci.* 27, 10073–10075. doi: 10.1523/JNEUROSCI.2747-07.2007

Rosenbaum, D. A. (2010). *Human Motor Control*. London: Academic Press.

Sailer, U., Flanagan, J. R., and Johansson, R. S. (2005). Eye-hand coordination during learning of a novel visuomotor task. *J. Neurosci.* 25, 8833–8842. doi: 10.1523/JNEUROSCI.2658-05.2005

Schneider, W., and Shiffrin, R. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* 84, 1–66. doi: 10.1037/0033-295X.84.1.1

Schneider, W. X. (1995). VAM: a neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. *Vis. cogn.* 2, 331–376. doi: 10.1080/13506289508401737

Schneider, W. X. (2013). Selective visual processing across competition episodes: a theory of task-driven visual attention and working memory. *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* 368, 1–13. doi: 10.1098/rstb.2013.0060

Schuck, N. W., Gaschler, R., and Frensch, P. A. (2012). Implicit learning of what comes when and where within a sequence: the time-course of acquiring serial position-item and item-item associations to represent serial order. *Adv. Cogn. Psychol.* 8, 83–97. doi: 10.2478/v10053-008-0106-0

Schützwohl, A. (1998). Surprise and schema strength. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 1182–1199. doi: 10.1037/0278-7393.24.5.1182

Schwarb, H., and Schumacher, E. H. (2012). Generalized lessons about sequence learning from the study of the serial reaction time task. *Adv. Cogn. Psychol.* 8, 165–178. doi: 10.2478/v10053-008-0113-1

Song, J.-H., and McPeek, R. M. (2009). Eye-hand coordination during target selection in a pop-out visual search. *J. Neurophysiol.* 102, 2681–2692. doi: 10.1152/jn.91352.2008

Võ, M. L.-H., and Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *J. Vis.* 9, 1–15. doi: 10.1167/9.3.24.Introduction

Võ, M. L.-H., Zwickel, J., and Schneider, W. X. (2010). Has someone moved my plate? The immediate and persistent effects of object location changes on gaze allocation during natural scene viewing. *Attent. Percept. Psychophys.* 72, 1251–1255. doi: 10.3758/APP.72.5.1251

Wischnewski, M., Belardinelli, A., Schneider, W. X., and Steil, J. J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cogn. Comput.* 2, 326–343. doi: 10.1007/s12559-010-9080-1

Conflict of Interest Statement: 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.

Copyright © 2016 Foerster. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

OPINION
published: 29 May 2017
doi: 10.3389/fpsyg.2017.00772

Violations of Expectations As Matter for the Believing Process

Hans-Ferdinand Angel[1]* *and Rüdiger J. Seitz*[2]

[1] Institute of Catechetics and Religious Pedagogics, University of Graz, Graz, Austria, [2] Department of Neurology, Centre for Neurology and Neuropsychiatry, LVR-Klinikum Düsseldorf, Medical Faculty, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

For the purpose of this communication it is postulated that violation of expectation means a disturbing event or conflict interfering with a previously established mental state that affords a firm belief or confident feeling. According to this hypothesis a violation of an expectation contradicts predictions and intentions that have been attained on stored experiences, valuations, and actual mood. We will argue that the notion of belief as static or stable which is usually described by expressions such as "my belief" or "our general belief" has to be extended to accommodate the process of belief formation. The credition model emphasizes the procedural aspect of belief by which the "process of believing" becomes similar to other psychological processes. We will describe that the "violation of expectation" can be decoded from the credition perspective and has brain functional correlates.

Keywords: credition, functional imaging, behavior, valuation, emotion, cognition

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Shihui Han, Peking University, China
Klaus Linde, Technische Universität München, Germany

***Correspondence:**
Hans-Ferdinand Angel
ferdinand.angel@uni-graz.at

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *05 October 2016*
Accepted: *26 April 2017*
Published: *29 May 2017*

Citation:
Angel H-F and Seitz RJ (2017) Violations of Expectations As Matter for the Believing Process. Front. Psychol. 8:772. doi: 10.3389/fpsyg.2017.00772

INTRODUCTION

In this paper we will argue that expectation for something to happen in the future is an important matter for the believing process. Consequently, we understand a violation of expectations as well as a matter for the believing process. More precisely, we want to introduce the idea that believing processes are underlying expectations as well as behavioral and neurophysiological reactions on their violations. To underpin this notion we will show that it is possible to describe reactions on such a (cognitive) conflict within the theoretical framework which is set by the novel model of credition. This approach can enrich the discussion about the violation of expectations by theoretical aspects which have not been discussed so far. Of critical relevance for this discussion is the model of credition which will be explained in some aspects in this contribution. Doing this involves multilevel data mapping (Paloutzian and Park, 2013; Paloutzian and Mukai, 2017) or bi-directionally "translating" the data and concepts from one level of analysis to an adjacent level of analysis in order to assess the degree to which they correspond. Specifically, this paper addresses violation of expectations at both a psychological and a neuroscientific level of analysis. The psychological level describes the mental processes involved in imagining, rendering beliefs out of a complex world of ambiguous information, and in the various verbal and conceptual puzzles created thereby. The neuroscientific level describes research on how these processes work in the human brain.

Thus, we want to discuss (1) a general hypothesis and (2) a specific hypothesis which is based on the general hypothesis.

> Hypothesis 1: Violation of expectation involves believing processes.
> Hypothesis 2: Within the model of credition "violation of expectation" can be expressed in relation to the so-called enclosure function.

To make understandable that "violation of expectation involves believing processes" we will show that it is possible to express "violation" in terms of credition (parts II–IV). To show that within the model of credition "violation of expectation" can be expressed in relation to the enclosure function we want to work out this aspect by translating a given example into a credition related language (parts V–VII). This intention requires a step by step presentation of the constituting features of the model of credition.

THE BELIEVING PROCESS – A NEW PERSPECTIVE ON VIOLATIONS OF EXPECTATIONS

It is uncommon to talk about the believing process. Rather, we are familiar with the use of the expressions "belief" or "faith". This use of nouns is widespread and predominant. Using nouns is not without effect as it insinuates (at least implicitly) the notion of belief as a state (Churchland and Churchland, 2013). Such stable beliefs have been found to follow a digital code, which is either true or false (Johnson et al., 2015). But, assuming the believing process to refer to mental activity or processes, it is more appropriate to apply the verbal expression "to believe". What on the first glance may give the impression of mere linguistic styling is, however, on the contrary a not trivial shift of understanding. This approach to the question of belief affords on several levels a change of thinking which can be labeled as "From the question of belief to the question of believing". Some aspects of this transformation have been explained elsewhere (Angel, 2017). But, for this communication we do not only refer to the novel concept of the believing process. We ground our reflections on "violation of expectation" on our model which seeks to simulate the psychological process of believing. This model will be the guideline for our perspective on the matter of violation of expectation.

The conceptual framework of the believing process and the hereof resulting "model of credition" assume that the believing process is a fundamental brain function that happens many times per day in everyday live (Angel and Seitz, 2016). The *model of the process of believing* includes a number of operational subfunctions that show surprising homology to neurophysiological processes as was described in detail recently. Central to the model is the so-called enclosure function which denotes the self-organizing probabilistic assembly of attributes of a given object or event into a coherent mental representation. These coherent knowledge constructs comprehend formal descriptions of the perceived encounters that can be expressed in terms of objective metrics as well as personal values associated with them as described below. Importantly, people employ these mental constructs for selecting the most appropriate for the subject in a given situation. In other words, perception is converted by the so-called converter function into an intended action which is part of and directed within an entire space of action. This cybernetic model assumes that the mental operations are mediated by a presumed operator in the human brain and can be stabilized by repetitions similarly to a learning process. Attitudes, hormonal states and pharmaceutical agents can modulate these mental operations.

Accordingly, processes of believing link the past sensory experience of a subject with his/her predictions for the future. These predictions correspond to personal expectations having emotional loadings of high subjectivity. The mental representations of the past experience are probabilistic in nature involving the attribution of subjective meaning to the perceptions (Seitz and Angel, 2014). Conversely, based on such probabilistic representations of the past, future acts are generated that are guided by probabilistic predictions of reward and cost to achieve a given goal (Barsalou, 2009; Angel and Seitz, 2016). As people act in their social environment they are constantly confronted with unexpected events and contradictions by others. In other words, humans experience a violation of their expectations all the time. Accordingly, violation of expectation is a frequent event that subjects need to be able to cope with. In its ultimate form a violation can result in a complete negation of an expectation. In this case it will lead to a heavy emotional challenge in the expecting subject influencing his/her subsequent behavior. Thus, there is good reason to assume that humans have to learn to cope with violation of their expectations. Such violations of expectations are defined events, while in contrast the probabilistic representations of meaning making and expectation have evolved over time by repetitive exposure and behavior. Thus, a violation of an expectation can – but does not necessarily – lead to a modification of the probabilistic representation or a certain belief.

In addition, subjects value objects and events in the outside world in terms of personal relevance (Seitz et al., 2009). These value judgments include introspection, goal values, decision values, and prediction errors (Hare et al., 2008). Here, we would like to define valuation as the process by which objects and events are evaluated by acting subjects in terms of utility and benefits. The probabilistic judgment is the default first person perspective of "what does it mean to me?" (Seitz and Angel, 2014). The judgment is loaded implicitly with emotional categories such as happiness, anger, fear, surprise, and disgust. These emotions induce immediate reactions of the subject and typically induce sensations from the inside of the body including raised and strong heartbeat, trembling, and heat in the head as was argued by Damasio (1998). The personal judgment involves automatic emotional processing as well as controlled cognitive processes as shown behaviorally and using event related recordings (Morewedge and Kahneman, 2010; Leuthold et al., 2015). These processes can eventually become consciously accessible to the subject being critical for guiding the subject's behavior.

Common to these cognitive processes is the relation to subjective categories such as memories, attitudes, desires, and hope (Corlett et al., 2004; Seitz et al., 2016). But these subjective categories can also be abstract categories of general value such as moral, justice, and ethics. Value judgments based on subjective perspective-taking are intimately linked to self-awareness which includes self-esteem, self-other distinctions, and the distinctiveness of one's own thoughts (Young and Pigott, 1999; Gallagher, 2000). Thereby, people experience themselves as causal agents and authors for their own actions and behavior resulting from a *post hoc* construction of an unconscious decision-making process (Gallagher, 2000; Wegner, 2003). Importantly, subjects judge the credibility of their inferences and predictions in terms of trustworthiness, convincingness, and substantiating evidence. In the positive case the subject arrives at the conviction that he/she accepts this personal interpretation as true or granted and, thus, personally relevant. Consequently, the subject believes it, since or although he/she does not know whether the information is really true (Seitz et al., 2016). Also, the emotional loading is part of the probabilistic mental representation of objects or events determining their relevance for the subject and the expectation the subject has concerning them. Ultimately, this can be translated to the realm of moral and ethics applying in groups and societies (Seitz et al., 2016). Accordingly, a violation of such an expectation is an emotional violation which will heavily affect the given subject's attitude what to learn already during infantile development (Stahl and Feigenson, 2015). Similarly, extinction learning has been shown to be able to profoundly influence behavioral patterns as in anxiety disorders (Pittig et al., 2016).

THE BELIEVING PROCESS – IMPLICIT CHANGES OF FOR THE NEW PERSPECTIVE

To understand "violation of expectation" from the perspective of the believing process we will describe explicitly the underpinnings of this innovative perspective. On the way from the question of belief to the question of believing we are elaborating here three aspects which are fundamental for the transformation of the traditional belief-related thinking.

Credition: Noun to Verb

It is a huge shift of paradigm to transform the noun-related concepts of "belief" into verb-related concepts. The focus on the topic of the process of believing can be expressed more precisely by the notion "while someone is believing".

Credition: Process

The mental activities underlying believing we encompass by the term credition. Importantly, they are to be understood as processes. This raises the question "what is a process". Here, we touch upon a long history of European thinking which has one of its excellent starting points in an understanding of the world as "fluent" which was brilliantly expressed by "πάντα ῥεῖ" (*panta rhei*, everything flows) as ascribed to the pre-Socratic Heraclitus. Also in modern philosophy there is a vivid discussion about the epistemic state of process thinking. This term was developed as a broad field of interest – it is controversial whether one should speak about process philosophy – and spawned in the writings of Bergson, Merleau-Ponty, and Whitehead, indicating that process constitutes change and occurs through and interacts with time. Time again is a highly controversial concept in philosophy and the understanding of time cannot be reduced to the matter of "measuring" time. We propose that to describe normal believing processes there is a need for a process-theoretical foundation (Angel, 2015). To transform noun-related concepts which understand belief in a static sense into a time-related concept of fluid processes of believing affords to bear on process theoretical concepts. Thereafter, the task of exploring to what extent the structure of credition is compatible with Whitehead's Metaphysics of Experience may be undertaken (cf. Maaßen, 2017).

Credition: Not on Religion

Finally, it has to be stated explicitly that the concept of credition is not located in the frame of religion. In fact, we want to stress that credition is not understood as a "religious" process. It is important to mention this as there is often a spontaneous association of religion with "belief/to believe". This connection between faith and religion has been coined by a long tradition of Western thinking. However, under a procedural perspective this connection is misleading. Importantly, credition applies to religious and secular contexts and it is not a prerequisite to refer to religion in order to understand credition (Seitz et al., 2016).

MODEL OF CREDITION – BASIC ASPECTS

Until shortly, there has been no term for the "believing process" that encompasses the notion in common language as well as in philosophy or cognitive science. To address this terminological challenge which hindered the interdisciplinary discourse the term "credition" was introduced into the scientific discussion (Angel, 2013a). The concept of "credition" originated from an anthropological view on religious experiences and consecutively from the attempt to understand "religiosity" (Angel, 2013b). Notably, the neologism "credition" was coined to denote believing processes that encompass both religious and secular beliefs. The term is derived from the Latin "*credere*" (to believe) and is shaped in analogy to other psychological terms like cognition (lat: *cogitare* = to think/to reflect) or emotion (lat. *movere* = to move).

The concept of credition claims that normal believing is inextricably interrelated with cognition and emotion (Sugiura et al., 2015; Angel, 2016). That brings the question on the floor how we can conceive the interaction of credition with interdependent cognitive and emotional processes. The model of credition proposes that believing comprises neuropsychological functions that overlap but do not equal those in cognition and emotion (Angel and Seitz, 2016).

In order to express "violation of expectation" in terms of the credition model it is necessary to outline some basic features of the credition model. It has to be mentioned that for the purpose of this presentation we assume the model of credition as given though there is ongoing scientific research on the character of the believing process[1]. For the reason of this paper the model of credition is sufficiently stable as it is supported by many data of different fields of research. Further, we postulate that violation of expectation means a disturbing event interfering with a previously established mental state that has afforded a firm belief or confident feeling. It should be emphasized that the believing process which has resulted in a firm belief or confident feeling belongs to the past. In contrast, the probabilistic expectation based on the outcome of the believing process which pertains to the future is violated by a momentary event.

Bab and Bab-Blob-Configuration

In the credition model the hypothesized processes are brought about and act upon meta-theoretical units to which heuristic labels were assigned. For this purpose we describe in the following paragraphs (a) the term "bab" and in consequence derived concepts as there is blob, bab-blob-configuration, and "characteristics of a bab", and (b) the enclosure function which has been introduced as one of four supposed functions of credition. Notably, one cannot describe the enclosure function without referring to the characteristics of babs. Vice versa, any explanation of any property of the relevant bab or of the property of the bab-blob-configuration is meaningful for an understanding of the enclosure function. The terms "bab" and "blob" are novel and have not existed so far (at least not in the here proposed sense). Why was it necessary to introduce those new terms? Two main reasons are:

The first reason is that recent scientific findings change our view on the relation of emotion and cognition but have not influenced yet our everyday language. "Bab" is a term which reflects these findings. The second reason is that a basic unit for credition is needed and the term (and concept of) "bab" can be offered as such a basic unit.

Overlapping Procession of Emotion and Cognition

Emotions and cognitions are considered as two different domains covering separate and partially contradictory aspects of brain function. There is empirical evidence from neuroimaging findings that emotion and cognition are processed in overlapping areas of the lateral prefrontal cortex by which both can contribute to the control of thought and behavior (Gray et al., 2002). Moreover, current data provide converging evidence that working memory and bioelectric activity in lateral prefrontal cortex can be influenced by affective variables (Schaefer and Gray, 2007; Roux et al., 2012). While emotions have been shown to involve the amygdala and the orbitofrontal cortex (Rolls, 2006), cognition comprises different aspects of mental activity such as speech production, memory processes, attention, and learning processes which are processed across widespread circuits in parietal, temporal, and frontal cortical areas as well as in the amygdala (Toga and Mazziotta, 2000; Schaefer and Gray, 2007). Beliefs are important to consider, as they were shown to influence reasoning and brain activity related to reasoning (Goel and Dolan, 2003). A given proposition, therefore, can differ in its personal emotional meaning.

As the European languages do not provide a term to express the overlap of cognition and emotion in a meta-theoretical sense, there is a discrepancy between the capacity of actual language(s) and the actual state of knowledge showing the need to supplement the word pool with terms which can express those given facts. To implement the neuroscientific findings into the frame of linguistic possibilities the term "bab" has been proposed (Angel, 2013a; Angel and Seitz, 2016). The term "bab" indicates in a linguistic, not in a mathematical sense: *"proposition + emotion"*.

Bab as Basic Unit of the Believing Process

The model of creditions emphasizes the process character of believing and by this the fluidity of beliefs. One of the most crucial questions is how to define the basic unit of the believing process. It is important that such a unit accommodates two basic claims:

- First, it has to provide a theoretical frame which accounts for the fluidity of the believing process and which allows to integrate different scientific descriptions (physical, biological, neural, behavioral, and so on).
- Second, it has to provide the possibility to integrate cognitive and emotional processes under a common label.

The term "bab" complies with both demands and we propose this term for such a new umbrella-term which has the capacity to indicate the basic unit of credition (see Supplementary Material Box 1).

Having declared "bab" as basic unit we can describe different characteristics which we assign to a single "bab". The term "blob" is used to indicate a subliminal "bab". We will come back to the question of subliminal processing below when we discuss the enclosure process.

First, we have to draw the attention to the characteristics of a bab. Owing to its mental function four characteristics can be assigned to a bab – and consequently to every single bab in a bab blob configuration.

- The propositional content: a "bab" can be described as a proposition as for example: "I see something red" or "I fell something sharp". The proposition becomes explicit by statements such as: "I see this ball to be red" or "I feel this knife to be sharp".
- The emotional moment: for example, a red light may be perceived as beautiful, warm or attractive, whilst a sharp item may be unpleasant, harmful and, thus, frightening. Note, that the term "bab" comprises the subliminal emotional moment in addition to the propositional content. When this information is expressed verbally, the "bab" will reach explicit awareness both in the speaking and the listening subject.
- The sense of mightiness: the perspective of a subject on a "bab" is not limited to the valence of an emotion but also includes the intensity of the emotion which is reflected by

[1] http://credition.uni-graz.at/de/credition-basic-research/

the "sense of mightiness". Thus, this scaling of an emotion as strong or weak is inherent in the proposition of a "bab".
- The sense of certainty: this characteristic reflects the conviction of an individual that a "bab" reflects the property of an object or event. The same proposition of a bab can have a high degree of certainty while for others it is uncertain. For instance, "I see something red" or "I see something sharp" has a high degree of certainty in daylight but a low degree of certainty in faint light.

Notably, in a believing process "babs" do not "exist" as single "monades" but as composite "bab-configurations". Specifically, "babs" include physical attributes such as color and form and personal attributes such as subjective meaning and relevance. In fact, "babs" represent pieces of knowledge with emotional loadings which are assembled into coherent knowledge constructs, the so-called stabilized "bab-blob-configuration" (see Supplementary Material Box 2).

The Four Functions of Credition

As outlined in the credition model, the believing process consists of four conceptually successive – but nevertheless in reality heavily interwoven – mental functions which are called enclosure function, converter function, stabilizer function, and modulator function (see Supplementary Material Box 3). Notably, one can speak about "converter function" or "converter process" depending on the perspective, which one choses to apply. In the following sections we will explain some aspects of the enclosure function.

With regard to the limitation of space we do not discuss more extensively the other functions in this paper. Just to mention that the converter function means that perception is converted into an intended action which is part of and directed within a space of action. This process employs the prediction of cost and reward and the expectation of future events inherent in a belief (Angel and Seitz, 2016). This cybernetic model of credition assumes that the mental operations are mediated by a presumed operator in the human brain and can be stabilized by repetitions similarly to a learning process. Attitudes, hormonal states, pharmaceutical agents and physical threatening that act on the entire individual can severely influence or modulate these mental operations.

We will not discuss the stabilizer function which is relevant for integration of experiences and their integration into a broader balance-dependent meaning making structure. What we want to state is that these three functions are regarded as universally effective functions whereas the fourth function which is called modulator function is strictly bound to individuals.

The Enclosure Function

In addition to neuropsychological topics such as perception, action, valuation, and stabilization one of the subfunctions of the model of credition is the so-called enclosure function. It denotes the self-organizing probabilistic assembly of mental attributes. Thus, the enclosure function is a mental process constituting or modifying "bab-configurations" or – in other words – different features of an object or event which are linked to each other to determine their characteristics and value. Under this perspective bab-configurations are subsets of mind-sets which are activated when a process of believing starts (Angel, 2013a; Angel and Seitz, 2016). The coherent knowledge constructs comprehend formal descriptions of the perceived encounters that can be expressed in terms of objective metrics as well as personal values associated with them. The personal values reflect the meaning and relevance the object or event has for the given individual (Seitz and Angel, 2014). Note, that the psychological description of the mental processes involved in imagining, making beliefs out of a complex world of ambiguous information, and of the various verbal and conceptual puzzles created thereby goes beyond the topic of this paper. Therefore it is reasonable to assume a systems level which is composed of a number of different meaning making processes and allows for flexible rearrangements of different meanings over time (**Figure 1**).

As many stimuli do not reach our consciousness, we have to accommodate also the subliminal aspect (Teske, 2007) in the credition model. As mentioned, for a bab which remains subconsciously the artificial term "blob" was introduced. That is the reason why we should speak of a "bab-blob-configuration" rather than of a "bab-configuration". We suggest that effects of placebo or nocebo (Myers et al., 1987; Benedetti et al., 2006; Jensen et al., 2012) are prominent examples for accounting for such a believing process (Meissner, 2017).

The term "clum" indicates the irritating moment which is in debate during the enclosure process. The name enclosure process is derived from the function by which an irritating clum is "included" or not into a bab-blob-configuration. The inner process which takes part in the period of "open result" comes to its end when the clum will be integrated or not in a previously existing bab-blob-configuration. Among other aspects processes of valuation are influential. Therefore, the enclosure function is interconnected with processes of valuation.

The enclosure process challenges the so-far existing bab-blob-configuration. In course of this process previously acquired "knowledge" which is stored in the actual bab-blob-configuration will be adjusted according to novel external stimuli and inner value terms associated with them. This process of adjustment is related to the inner balance system as well as to the meaning-making system. The believing process serves to cope with homeostatic challenges. On a basic level we can see homeostatic bodily processes. Finally, we have to stress that a clum also is a "bab", but one with a specific property during the enclosure process.

"VIOLATION OF EXPECTATION" IN TERMS OF CREDITION

Based on the model of credition a "violation of an expectation" can be understood as a mental process which leads to the "*realization that a given bab-blob-configuration includes (or included) an inadequate bab.*" Within the framework of the credition model the specific characteristic of a "violation of expectation" is related to the so called "clum" which indicates an irritating moment. A "clum" has a crucial relevance for the so-called enclosure function and by this for the initiation of

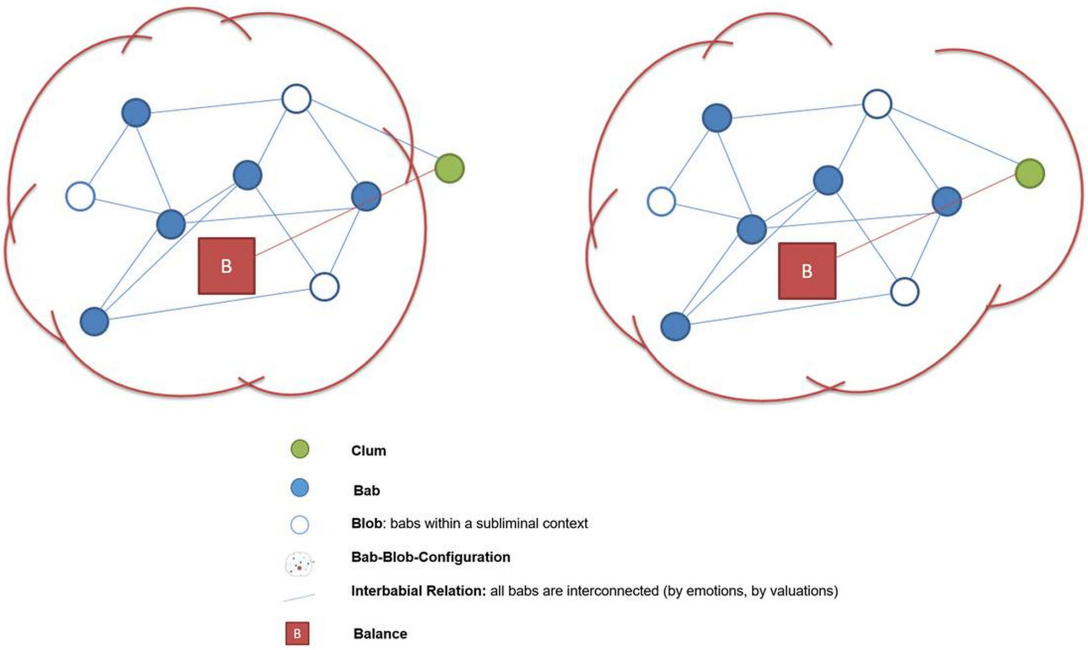


FIGURE 1 | The enclosure function involves the composition of identified (bab) and subliminal (blob) properties of an object or event. These compositions are balanced but in continuous evolution allowing the enclosure even of a property with opposite valence (clum).

a process of credition. But with regard to an expectation a "clum" must have a well-defined property. According to our understanding that "violation of an expectation" can be defined as "realization that a given bab-blob-configuration includes (or included) an inadequate bab" we can formulate the hypothesis: *the propositional content of a clum is identical with one of the babs in the agent's configuration but (mathematically spoken) with a negative algebraic sign.*

As an example for a "violation of expectation" in terms of credition we present the following situation. The example is that someone has booked a flight. Accordingly, the person believes that he/she will be in the position to travel to the desired destination and has engaged in the actions mandatory to prepare this trip. When approaching the gate the person expects to receive the boarding pass and to get on the plane. But then the person is confronted with the unpredicted information "the flight is overbooked".

Our following discussion refers mainly to the *characteristics of babs* as well as to the *enclosure function* of credition. Nevertheless, we want to draw the attention to the fact that our given explanation is not a comprehensive description of the enclosure function but will highlight only some of the indispensable aspects.

Irritation as Production of a Clum

The fact or event which violates an expectation has to be described as "irritating moment" and transformed in such a way that it can become a "clum". As mentioned, detecting an irritating moment is the normal precondition for any beginning of a believing process and the initiation of an enclosure function (cf. part III). In so far it is a matter of perception if something is detected as irritating signal. In our example the irritating moment "the flight is overbooked" is communicated as information in words and addresses the auditory sensory system. Of course the characters of signals and the mode of their perception can differ heavily. For instance, processing a perceived *static object* differs in several aspects from processing a perceived event which has to be coded temporally. But the differences related to the property of the perceived "irritation" do not change the general explanation of how a clum is integrated.

How can we conceive the above mentioned hypothesis that the propositional content of a clum is identical with one of the babs in the agent's configuration but (mathematically spoken) with a negative algebraic sign? For answering this question we have to explain what aspects can be ascribed to a clum in case of "violation of expectation". For this we have to clarify what might be the propositionally identical content of a clum and of a bab. Here, we have to acknowledge that the notion of "violation" can only be understood as a distinct event in time, while a belief pertains over time. This means that the concepts of believing and of violation accommodate different temporal aspects.

To understand the "character of the violation" we have to start at the moment when the "frame" for a possible violation was settled. In our example this is the moment when the booking of the flight was accomplished. After having booked the flight a person will have established a mental state that affords a firm belief or confident feeling that he or she will be able to use exactly this flight. We can translate the end of the booking operation like follows: the agent has included into his or her bab configuration a bab with the propositional content "the flight is available".

Connection of Cognition and Emotion

As a "bab" by definition is understood as "proposition plus emotion" (cf. part III) we assume that the emotional loading of this specific "bab" will be "joy". We cannot discuss here how the emotional loading (joy) interacts with the cognitive process which takes as rationally undoubtable that the "flight will be available".

Interbabial Relations

Nor can we discuss how the emotionally positive bab "flight will be available" interacts with other babs in the configuration. Of course, these configurations will differ for different subjects depending on their individual experiences. If someone never has come into such a situation he or she probably will not have included an emotionally mighty bab "flights are not guaranteed by booking". On the contrary, a frequent flyer will have integrated such a bab in his or her configuration.

Propositional Contradiction of the Clum and One of the Babs

Now imagine what will happen when the person gets the information: "the flight is overbooked". In order to be able to verify the hypothesis we have to check whether this information can be translated into a formulation which is identical with the propositional content of the bab "the flight is available". Under linguistic aspects the information "the flight is overbooked" is negatively identical with the propositional content of the bab "the flight is available". Thus, after getting the overbooking information we have the following situation:

- Bab in the bab configuration:
  "the flight is available" plus emotion "joy"
  [(+ proposition) × emotion(1)]
- Clum
  "the flight is not available" plus emotion "anger"
  [(− proposition) × emotion(2)].

As mentioned this formulation is understood linguistically, not mathematically. Mathematically, it should be written as product because the emotion does not come additionally to the proposition but simultaneously. Thus, the use of the term "bab" stresses this interconnectedness of propositional and emotional aspects. When the person "believes" that the flight is overbooked he or she has to integrate the clum into his previously established mental state. After the integration of this negatively loaded clum also the emotional value of his or her bab-blob-configuration will have been changed into a more negative set. Besides, the full integration of the clum into the bab-blob-configuration marks the end of the enclosure process.

Bab and Clum: Cognitive Dissonance

When regarding the content level we will observe a mental dynamic which is caused by the interaction of two contradicting babs. This kind of problem is described by the concept of cognitive dissonance (Festinger, 1957). In his influential cognitive dissonance theory, Festinger included believing in the class of dissonance reduction processes. Accordingly, believing is to change or to add a cognitive element to reduce dissonance with or between other cognitive elements. For example, the dissonance between two ideas, a belief that people are good in general, and a knowledge that children go through a period of aggressive behavior, is reduced by believing existence of malevolent ghosts which enter into children and cause them to do inappropriate things. The idea of dissonance reduction appears to fit well with the explanation of human brain function in the free-energy principle as an optimizing machinery for value and its counterpart surprise (Friston, 2010). Fundamental herein are the probabilistic predictions of value or reward concerning perceived information and of expected error or cost concerning future actions, which drive the system to the next state by a simple principle of reducing the free energy. Believing is one of the conscious expressions of such a self-organizing process.

Bab and Clum: The Degree of Certainty

The degree of certainty of the bab "the flight is overbooked" may differ according to experience. Though everybody knows theoretically that "flights are never guaranteed by booking" an agent may act during further steps of decision making as if the bab in question has a high degree of certainty and not prepare a plan B while another agent may attribute a lower degree of certainty. In everyday language he or she might comment "one never knows". In terms of credition the degree of certainty influences the activity among the babs within the bab-blob-configuration. A lower degree of certainty will have as consequence a more fluid configuration which results in a higher flexibility of the agent.

Bab and Clum: Mismatch of Emotions

On the level of emotions we will have turbulences which are caused by the interference of two distinguished emotions – joy and anger. That brings up the question what happens with a bab whose propositional content has a double-loading with different emotion. How will the originally emotional loading "joy" be infected by an arising anger? Will the anger be raised due to the original joyful base or will it be generated spontaneously without recall of the original joyful state? Questions like these open the field for discussions of emotional interaction. Taking into account a "circumplex model of emotions" one can develop a differentiated view on emotions and assume that different emotions influence each other. One can discriminate primary and secondary emotions and assume families of emotions based on similarity (Plutchik, 2001).

Bab and Clum: The Mightiness of Emotions

Partly the mightiness of the emotional loading of the clum "flight is overbooked" will depend on the alternatives. If someone deplores to miss a marvelous concert due to the early flight the information that the flight is overbooked might stimulate as first reaction that there will be a chance now to visit the concert.

Bab and Clum: Match of Propositional Information and Emotions

But, from the perspective of credition the focus of interest will be on the question: how the turbulences can be described which

are caused by interference of emotion and information. Here the question is to be discussed whether and in which way emotions can be seen as information (Schwarz, 2001, 2011). Of course, the enclosure process is a question of energy. Partly, it is energy consuming and has to be observed under respect of free energy, partly it will set free energy which can be used for action (Friston, 2010).

Enclosure Function and Time for Integration the Clum

Another aspect is the question of time. How long does it take until the clum "the flight is not available" is incorporated? That is identical with the question of how long the enclosure process will take place. On a fundamental neurophysiological level this is an open question.

However, an important and unanswered question comes to the surface: is what constitutes the knowledge that is stored in the brain merely deposited at once as facts and information, or is it the result of processes (Krüger et al., 2009)? Strong arguments have been made for both views. Experiments on brain–computer interfaces provide good evidence that processes are among the things represented in the brain because, e.g., subjects can learn to actively modulate their brain activity in order to move their paralyzed arm or to write words and even sentences (Birbaumer, 2014).

Further Aspects

The model of credition provides a couple of further aspects which should be taken into account when describing the character of the possible interaction of a clum with a bab-blob-configuration. This would be for instance the influence of subliminal processes on the consciously perception of a violation. We suppose that these subliminal effects which can be described on neuropharmacological (Holzer, 2017) and microbiological (Sensen and Berg, 2017) levels have to be taken into account in a much broader sense than we have been used to acknowledge. Or, to give a second example, the role of the characteristics of babs should be discussed more deeply with the violation process. This would give deeper insights into the effects which result from a change of emotional mightiness (from mega-bab to mini-bab or inverse). In a similar manner it should be reflected how a modification of the degree of certainty has to be understood – as a sudden event or as an act which is going to happen in a creeping way (Huber and Schmidt-Petri, 2009). Or, again another point: a broader discussion would be needed about how we can understand the interaction of the mere biological homeostatic balance system with the higher level (quasi-homeostatic) system of meaning-making. But those aspects we have to omit with regard to space.

HIDDEN POLYVALENCE OF THE NOTION "VIOLATION OF EXPECTATION"

It might be trend-setting to identify believing as a crucial process which influences the development of expectations as well as the handling of their violations. This will allow us to conceive "expectation" as a (preliminary) stabilized state resulting from continuously running believing processes. "Violation of expectation" can be interpreted as an event which reopens the next turn of believing processes that end with the final integration of the violating clum into the reorganized bab-blob-configuration. Using the perspective of the model of credition we can state that the expression "violation of expectation" is an umbrella term which covers a wide range of possible notions. The model of credition allows us to understand the semantic ambiguity which is inherent in the notion "violation of expectation". We will explain this view with a few examples of decoding possibilities afforded by the credition model.

"Violation of expectation" can be decoded as:

Change of the Emotional Shape of a Bab Configuration after Integration of the Clum

In our example the clum "flight overbooked" probably will be combined with negative emotion like anger. When the enclosure process comes to its end the negatively shaped proposition "flight overbooked" will be integrated into the former bab-blob-configuration. This of course will influence the emotional shape of the entire bab-blob-configuration and it can be observed how the emotional loading of the clum will influence in course of the time the configuration. This process can be conceived as a coping process. On a psychological level we will find as a result the modified mood of the person.

Obscuring the Space of Action and Hindering Decision Making

One can interpret the integration of the clum with regard of the converter process (which is relevant for the configuration of the space of action). In this case is relevant that an integrated clum will destabilize the existing bab-blob-configuration. As a result we will see a modification of the impulses which are relevant for action. It will be less clear "in which direction" the space of action will be opened. As the space of action is understood as the preliminary state of a decision, the ambiguity of impulses can be understood as an obstacle for a quick decision.

Destabilization of the Balance System

In case of great importance of the previous integration of the bab "flight available" the clum "flight not available" can have strong consequences. Depending on the emotional mightiness of the clum the integration can touch heavily on the balance system. We can easily imagine the case that the flight was booked to visit a beloved person of poor health. The need to integrate the clum "flight not available" might touch the traveler's balance system and provoke serious bodily reactions.

Reopening of the Believing Process

In the case that the destabilization of bab-blob-configuration is detected and perceived as an irritation a next turn of the believing process can commence. This does not predict in which direction the space of action will be opened. It will definitely

be different when the now upcoming clum has the proposition "change needed" or the proposition "not with me!"

Finally, we want to mention that it will be possible to interpret standard positions toward "violation of expectation" in the light of the model of credition. Using the language of credition it will be possible to assign the concept of the believing process to existing models of expectation. For instance Bandura's concept of self-efficacy can be interpreted in terms of credition as "existence of a so-called mega-bab with the properties: [a] the proposition/content "I am efficient", [b] positive/joyful [c] emotional loading by which the [d] degree of certainty of the proposition is augmented. When trying to translate Pavlov's concept of conditioned reflexes into a language of credition we would focus more on the relation of the modulator function and the stabilizer function.

NEUROPHYSIOLOGICAL FOUNDATION OF PROCESSES OF BELIEVING AND THEIR VIOLATION

Functional magnetic resonance imaging (MRI) is suited to identify the areas involved in the working human brain. As we have outlined above, the believing process is an integrative brain function involving a number of psychophysical subfunctions. Here, we are outlining some recent empirical data about the implementation of such integrative functions in the human brain. Most information is in the subliminal or preconscious domain but, nevertheless involves the activation of extensive cerebral networks including the lateral prefrontal cortex (Changeux and Dehaene, 2008). In particular, gamma-oscillations have been advanced as a candidate functional expression for binding information of different origins into a coherent representation in working memory (Roux et al., 2012). The global workspace integrating perception and valuation and allowing for generation of appropriate action is critically modified by previous experience and by the momentary focus of attention (Koechlin and Summerfield, 2007; Mesulam, 2008; Dehaene and Changeux, 2011). In this process identification of conflict – the violation of expectation – is of fundamental importance. From a large body of evidence in the open literature we know that the anterior cingulate is a critical node in processing conflict (Carter and van Veen, 2007). A further important field of interest with relevance to the discussion in this paper is the generation and inhibition of behavior. This is due to the fact that a violation of expectation influences the individuals' behavior by affecting their prospects of long- and short-term reward. MRI studies showed that normal preparatory activity in the premotor and posterior parietal cortex can be modulated by the subjective absolute value (in terms of monetary consequences) of an upcoming action (Iver et al., 2010). Specifically, subjects who had large gains and believed they performed well, and subjects who had large losses and believed they performed poorly, had the highest preparatory signals. The neural activity in the medial frontal gyrus appears to link unexpected sensory information including violation of reward prediction (Martin et al., 2009; Schwartenbeck et al., 2016) with preparatory control of arm movements but also affording response inhibition and task switching (Rushworth et al., 2002; Leung and Cai, 2007; Chen et al., 2010). In particular, the supplementary motor area (SMA) was shown to be involved in free choice movement coding (Nachev et al., 2008; Passingham et al., 2010; Pfurtscheller et al., 2014). The SMA and premotor areas are also involved in judgment of aesthetics as well as brightness, which signifies that the SMA has more general behavioral relevance (Ishizu and Zeki, 2011, 2013). Conversely, a number of distinct nodes in the medial frontal cortex, including the SMA and pre-SMA, are involved in the proactive and inhibitory control of actions (Seitz et al., 2006; Van Overwalle, 2009).

In addition to cortical brain areas, such an integration of this different type of information was shown to take place by involvement of the basal ganglia. There is evidence from rat T-maze experiments that activities modulated to different frequencies can develop in parallel in different subregions of the striatum, allowing for a coordinated flow of information through different *trans*-striatal networks and, thereby, for simultaneous and independent operations in separate networks (Thorn and Graybiel, 2014). Furthermore, the modulation of cortical information by processing in *trans*-striatal relay loops has been described as of key importance for learning routines and rules as well as their combinations (Graybiel and Grafton, 2015). Recently, it was shown that shifts in beliefs involve dopamine-rich midbrain regions (Schwartenbeck et al., 2016).

Since the individuals' capacity to deal with on-line information is limited (Baddeley, 1981), probabilistic representations and predictions assist the person to arrive at behavioral decisions. This is because beliefs can be envisioned to guide the individual's choices, although they limit his/her space of action. MRI studies showed that preparatory activity in the premotor and posterior parietal cortex is modulated by the subjective absolute value of an upcoming action (Iver et al., 2010). A compelling argument for the relevance of functional neuroanatomic data comes from neurological patients showing that a given neuropsychic function is impaired due to damage to a certain brain structure that is involved in executing this function in healthy volunteers. A large meta-analysis of 193 studies showed that a loss of gray matter in brain structures belonging to the salience network, including the anterior insula and dorsal anterior cingulate, was related to deficits in executive functions in patients with different mental illnesses (Goodkind et al., 2015).

Studies of this sort show that the brain structures mediating adequate behavior in healthy subjects are compromised in mental illnesses. Although there is no causal link, it is likely that the integrative brain functions such a meaning making, prediction of future events, control of behavior and realizing of a violation of expectation are impaired in such patients. For example, patients with delusions have severe deficits in mental processing of perception, memory, bodily agency, social learning and are, thus also impaired in predicting future events in the external world (Corlett et al., 2010). Likewise, neuroimaging studies in

psychopaths have shown that these persons are impaired in increasing activity in the anterior insula (Sitaram et al., 2014) which was paralleled by lower conditioned fear responses (Veit et al., 2013). In addition, the so-called alien limb syndrome which represents a violation of the sense of body integrity has been related to damage of the parietal cortex (Graff-Radford et al., 2013) and the medial frontal cortex (Feinberg et al., 1992; Biran et al., 2006) the latter of which was also related to an abolished self-reference (Philippi et al., 2012). Evidence from functional imaging studies has revealed that the medial fronto-parietal circuit is critically abnormal in post-traumatic distress disorder reflecting altered mental functioning secondarily to a profound violation of the sense of safety (Cwik et al., 2014). In fact, important aspects of believing, such as personal reference, empathy, and adequate control of behavior, appear to rely on the integrity of the medial and lateral prefrontal cortex. Adequate control of behavior means resistance to react in case of violation of expectation, which is possible even with a low time limit of the cueing and/or go-signal of about 200 ms (Schultze-Kraft et al., 2016).

Limitations

By this paper we hope to contribute to a more comprehensive understanding of the complex interaction of violation of expectation and the process of believing. This can be interpreted as a severe conflict of prediction error with previous experience. To the best of our knowledge we do not know of any other model of the believing process. We would like to open a new field of discussion as beliefs and the believing process appear as "possible targets for neuroscientific research" (Seitz, 2017). Our discussion here reflects mainly the question of how and to which extent previous and current, in principle, static approaches to the question of belief can contribute to our understanding of the *process* character of belief formation.

There are, however, limitations which are caused by the need to present the believing process and the functions of credition in a condensed manner. A less abbreviated presentation could and should explain many aspects much more in detail.

First, we did not discuss here the whole range of possible aspects. Thus, we omitted for instance the developmental aspect which should be reflected for children, aging persons, and so on. We did not discuss aspects concerning the impacts of traumas on violations of expectation, or more generally the topic of coping as for instance "learned helplessness" (Abramson et al., 1978). Nor did we expand on violation of expectation under the perspective of neuro- or psychopathy, which may be caused by a disturbance of balance (Devinsky, 2009). Moreover, we did not extend the reflection toward other cultural areas.

There are also theoretical limitations which depend on the actual state of research and the available neurophysiological data. There are general limitations which are partly connected with philosophical presumptions. In this regard there are specific limitations which depend on the hermeneutic question of translation the model of credition into an everyday language as well as into a scientifically adequate expression. These limitations may challenge young researchers of different disciplines like philosophy (epistemology, philosophy of mind), psychology, neurology, or with interest in different relevant fields like conflict solving, leadership, or mediation.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The Credition Project was supported by the City of Graz, Austria.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00772/full#supplementary-material

REFERENCES

Abramson, L. Y., Seligman, M. E. P., and Teasdale, J. D. (1978). Learned helplessness in humans: critique and reformulation. *J. Abnorm. Psychol.* 87, 49–74. doi: 10.1037/0021-843X.87.1.49

Angel, H. F. (2013a). "Credition", in *Encyclopedia of Sciences and Religion*, eds A. L. C. Runehov, L. Oviedo, and N. P. Azari (Dordrecht: Springer), 536–539. doi: 10.1007/978-1-4020-8265-8_1565

Angel, H. F. (2013b). "Religiosity", in *Encyclopedia of Sciences and Religion*, eds A. L. C. Runehov, L. Oviedo, and N. P. Azari (Dordrecht: Springer), 2012–2014. doi: 10.1007/978-1-4020-8265-8_1503

Angel, H. F. (2015). "Process and creditions: How to understand the process of believing?", in *Advances in Process Thought: Society, Education, and God*, eds J. Dziadkowiec and L. Lamza (Newcastle upon Tyne: Cambridge Scholars Publishing), 265–280.

Angel, H. F. (2016). No believing without emotion: the overlapping of emotion and cognition in the model of credition. *Stud. Sci. Theol.* 15, 215–222.

Angel, H. F. (2017). "Credition: from the question of belief to the question of believing", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H. F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 17–36. doi: 10.1007/978-3-319-50924-2_2

Angel, H. F., and Seitz, R. J. (2016). Processes of believing as fundamental brain function: the concept of credition. *SFU Res. Bull.* 1, 1–20. doi: 10.15135/16.4.1.1-20

Baddeley, A. (1981). The concept of working memory: a view of its current state and probable future development. *Cognition* 10, 17–23. doi: 10.1016/0010-0277(81)90020-2

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1281–1289. doi: 10.1098/rstb.2008.0319

Benedetti, F., Amanzio, M., Vighetti, S., and Asteggiano, G. (2006). The biochemical and neuroendocrine bases of the hyperalgesic nocebo effect. *J. Neurosci.* 26, 12014–12022. doi: 10.1523/JNEUROSCI.2947-06.2006

Biran, I., Giovannetti, T., Buxbaum, L., and Chatteriee, A. (2006). The alien hand syndrome: What makes the alien hand alien? *Cogn. Neuropsychol.* 23, 563–582. doi: 10.1080/02643290500180282

Birbaumer, N. (2014). Neural signatures of modified memories. *Neuron* 81, 3–5. doi: 10.1016/j.neuron.2013.12.019

Carter, C. S., and van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cogn. Affect. Behav. Neurosci.* 7, 367–379. doi: 10.3758/CABN.7.4.367

Changeux, J.-P., and Dehaene, S. (2008). "The neuronal workspace model: conscious processing and learning", in *Learning Theory and Behavior*, Vol. 1, eds J.-P. Changeux and R. Menzel (Oxford: Elsevier), 729–757.

Chen, X., Scango, K. W., and Stuphorn, V. (2010). Supplementary motor area exerts proactive and reactive control of arm movements. *J. Neurosci.* 30, 14657–14675. doi: 10.1523/JNEUROSCI.2669-10.2010

Churchland, P. S., and Churchland, P. M. (2013). "What are beliefs?", in *The Neural Basis of Human Belief Systems*, eds F. Krueger and J. Grafman (Hove: Psychology Press), 1–17.

Corlett, P. R., Aitken, M. R., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A., et al. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron* 44, 877–888. doi: 10.1016/j.neuron.2004.11.022

Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C., and Krystal, J. H. (2010). Toward a neurobiology of delusions. *Prog. Neurobiol.* 92, 345–369. doi: 10.1016/j.pneurobio.2010.06.007

Cwik, J. C., Sartory, G., Schürholt, B., Knuppertz, H., and Seitz, R. J. (2014). Posterior midline activation during symptom provocation in acute stress disorder: an fMRI study. *Front. Psychiatry* 5:49. doi: 10.3389/fpsyt.2014.00049

Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. *Brain Res. Rev.* 26, 83–86. doi: 10.1016/S0165-0173(97)00064-7

Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Devinsky, O. (2009). Delusional misidentifications and duplications Right brain lesions, left brain delusions. *Neurology* 72, 80–87. doi: 10.1212/01.wnl.0000338625.47892.74

Feinberg, T. E., Schindler, R. J., Flanagan, N. G., and Haber, L. D. (1992). Two alien hand syndromes. *Neurology* 42, 19–24. doi: 10.1212/WNL.42.1.19

Festinger, L. A. (1957). *Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/Nrn2787

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Goel, V., and Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition* 87, B11–B22. doi: 10.1016/S0010-0277(02)00185-3

Goodkind, M., Eickhoff, S. B., Oathes, D. J., Jiang, Y., Chang, A., Jones-Hagata, L. B., et al. (2015). Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry* 71, 1222–1230. doi: 10.1001/jamapsychiatry.2014.1100

Graff-Radford, J., Rubin, M. N., Jones, D. T., Aksamit, A. J., Ahlskog, J. E., Knopman, D. S., et al. (2013). The alien limb phenomenon. *J. Neurol.* 260, 1880–1888. doi: 10.1007/s00415-013-6898-y

Gray, J. R., Braver, T. S., and Raichle, M. E. (2002). Integration of emotion and cognition in the lateral prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4115–4120. doi: 10.1073/pnas.062381899

Graybiel, A. M., and Grafton, S. T. (2015). The striatum: where skills and habits meet. *Cold Spring Harb. Perspect. Biol.* 7:a021691. doi: 10.1101/cshperspect.a021691

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociation the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction error. *J. Neurosci.* 28, 5623–5630. doi: 10.1523/JNEUROSCI.1309-08.2008

Holzer, P. (2017). "Interoception and gut feelings: unconscious body signals' impact on brain function, behavior, and belief processes", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H. F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 435–442. doi: 10.1007/978-3-319-50924-2_31

Huber, F., and Schmidt-Petri, C. (2009). *Degrees of Belief*. Heidelberg: Springer. doi: 10.1007/978-1-4020-9198-8

Ishizu, T., and Zeki, S. (2011). Toward a brain-based theory of beauty. *PLoS ONE* 6:e21852. doi: 10.1371/journal.pone.002/1852

Ishizu, T., and Zeki, S. (2013). The brain's specialized systems for aesthetic and perceptual judgment. *Eur. J. Neurosci.* 37, 1413–1420. doi: 10.1111/ejn.12135

Iver, A., Lindner, A., Kagan, I., and Andersen, R. A. (2010). Motor preparatory activity in posterior parietal cortex is modulated by subjective absolute value. *PLoS Biol.* 8:e1000444. doi: 10.137/journal.pbio.1000444

Jensen, K. B., Kaptchuk, T. J., Kirsch, I., Raicek, J., Lindstrom, K. M., Berna, C., et al. (2012). Nonconscious activation of placebo and nocebo pain responses. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15959–15964. doi: 10.1073/pnas.1202056109

Johnson, S. G. B., Merchant, T., and Keil, F. C. (2015). "Predictions from uncertain beliefs", in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, eds D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, et al. (Austin, TX: Cognitive Science Society), 1003–1008.

Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci.* 11, 229–235. doi: 10.1016/j.tics.2007.04.005

Krüger, F., Barbey, A. K., and Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends Cogn. Sci.* 13, 103–109. doi: 10.1016/j.tics.2008.12.005

Leung, H. C., and Cai, W. (2007). Common and differential ventrolateral prefrontal activity during inhibition of hand and eye movements. *J. Neurosci.* 7, 9893–9900. doi: 10.1523/JNEUROSCI.2837-07.2007

Leuthold, H., Kunkel, A., Mackenzie, I. G., and Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. *Soc. Cogn. Affect. Neurosci.* 10, 1021–1029. doi: 10.1093/scan/nsu151

Maaßen, H. (2017). "The structure of credition in whitehead's metaphysics of experience", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H. F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 217–236. doi: 10.1007/978-3-319-50924-2_16

Martin, L. E., Potts, G. F., Burton, P. C., and Montague, P. R. (2009). Electrophysiological and hemodynamic responses to reward prediction violation. *Neuroreport* 20, 1140–1143. doi: 10.1097/WNR.0b013e32832f0dca

Meissner, K. (2017). "Believing in the effectiveness of treatment: from placebo to credition and back", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H. F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 125–137. doi: 10.1007/978-3-319-50924-2_9

Mesulam, M. (2008). Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. *Ann. Neurol.* 64, 367–378. doi: 10.1002/ana.21534

Morewedge, C. K., and Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends Cogn. Sci.* 14, 435–440. doi: 10.1016/j.tics.2010.07.004

Myers, M. G., Cairns, J. A., and Singer, J. (1987). The consent form as a possible cause of side effects. *Clin. Pharmacol. Ther.* 42, 250–253. doi: 10.1038/clpt.1987.142

Nachev, P., Kennard, C., and Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nat. Rev. Neurosci.* 9, 856–869. doi: 10.1038/nrn2478

Paloutzian, R. F., and Mukai, K. J. (2017). "Believing, remembering, and imagining: the roots and fruits of meanings made and remade", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H.-F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 39–49.

Paloutzian, R. F., and Park, C. L. (2013). "Directions for the future of psychology of religion and spirituality: research advances in methodology and meaning systems", in *Handbook of the Psychology of Religion and Spirituality*, 2nd Edn, eds R. F. Paloutzian and C. L. Park (New York City, NY: Guilford Press), 651–665.

Passingham, R. E., Bengtsson, S. L., and Lau, H. C. (2010). Medial frontal cortex: from self-generated action to reflection on one's own performance. *Trends Cogn. Sci.* 14, 16–21. doi: 10.1016/j.tics.2009.11.001

Pfurtscheller, G., Andrade, A., Koschutnig, K., Brunner, C., and Lopes da Silva, F. (2014). Initiation of voluntary movements at free will and ongoing 0.1 Hz BOLD oscillations in the insula – a pilot study. *Front. Integr. Neurosci.* 8:93. doi: 10.3389/fnint.2014.00093

Philippi, C. L., Duff, M. C., Denburg, N. L., Tranel, D., and Rudrauf, D. (2012). Medial PFC damage abolishes the self-reference effect. *J. Cogn. Neurosci.* 24, 475–481. doi: 10.1162/jocn_a_00138

Pittig, A., van den Berg, L., and Vervliet, B. (2016). The key role of extinction learning in anxiety disorders: behavioral strategies to enhance exposure-based treatments. *Curr. Opin. Psychiatry* 29, 39–47. doi: 10.1097/YCO.0000000000000220

Plutchik, R. (2001). The nature of emotions. *Am. Sci.* 89, 344–350. doi: 10.1511/2001.4.344

Rolls, E. T. (2006). Brain mechanisms underlying flavour and appetite. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 1123–1136. doi: 10.1098/rstb.2006.1852

Roux, F., Wibral, M., Mohr, H. M., Singer, W., and Uhlhaas, P. J. (2012). Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. *J. Neurosci.* 32, 12411–12420. doi: 10.1523/JNEUROSCI.0421-12.2012

Rushworth, M. F. S., Hadland, K. A., Paus, T., and Sipila, P. K. (2002). Role of the human medial frontal cortex in task switching: a combined fMRI and TMS study. *J. Neurophysiol.* 87, 2577–2592. doi: 10.1152/jn.00812.2001

Schaefer, A., and Gray, J. R. (2007). A role for the human amygdala in higher cognition. *Rev. Neurosci.* 18, 355–363. doi: 10.1515/REVNEURO.2007.18.5.355

Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., et al. (2016). The point of no return in vetoing self-initiated movements. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1080–1108. doi: 10.1073/pnas.1513569112

Schwartenbeck, P., FitzGerald, T. H. B., and Dolan, R. (2016). Neural signals encoding shifts in beliefs. *Neuroimage* 125, 578–586. doi: 10.1016/j.neuroimage.2015.10.067

Schwarz, N. (2001). "Feelings as information. Implication for affective influences on information processing", in *Theories of Mood and Cognition. A User's Guidebook*, eds L. L. Martin and G. L. Clore (London: Psychology Press), 159–176.

Schwarz, N. (2011). "Feeling-as-Information theory", in *Handbook of Theories of Social Psychology*, eds P. Van Lange, A. Kruglanski, and E. T. Higgins (London: Sage Publishing).

Seitz, R. J. (2017). "Beliefs and believing as possible targets for neuroscientific research", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H.-F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 69–80. doi: 10.1007/978-3-319-50924-2_5

Seitz, R. J., and Angel, H. F. (2014). Psychology of religion and spirituality: meaning making and processes of believing. *Relig. Brain Behav.* 5, 118–178. doi: 10.1080/2153599X.2014.891249

Seitz, R. J., Franz, M., and Azari, N. P. (2009). Value judgments and self-control of action: the role of the medial frontal cortex. *Brain Res. Rev.* 60, 368–378. doi: 10.1016/j.brainresrev.2009.02.003

Seitz, R. J., Nickel, J., and Azari, N. P. (2006). Functional modularity of the medial prefrontal cortex: involvement in human empathy. *Neuropsychology* 20, 743–751. doi: 10.1037/0894-4105.20.6.743

Seitz, R. J., Paloutzian, R. F., and Angel, H.-F. (2016). Processes of believing: Where do they come from? What are they good for? *F1000Research* 5:2573. doi: 10.12688/f1000research.9773.2

Sensen, M., and Berg, G. (2017). "Decision-making and credition under a microbial perspective", in *Process of Believing: The Acquisition, Maintenance, and Change in Creditions*, eds H. F. Angel, L. Oviedo, R. F. Paloutzian, A. L. C. Runehov, and R. J. Seitz (Dordrecht: Springer), 443–450.

Sitaram, R., Caria, A., Veit, R., Gaber, T., Ruiz, S., and Birbaumer, N. (2014). Volitional control of the anterior insula in criminal psychopaths using real-time fMRI neurofeedback: a pilot study. *Front. Behav. Neurosci.* 8:344. doi: 10.3389/fnbeh.2014.00344

Stahl, A. E., and Feigenson, L. (2015). Cognitive development. Observing the unexpected enhances infants' learning and exploration. *Science* 348, 91–94. doi: 10.1126/science.aaa3799

Sugiura, M., Seitz, R. J., and Angel, H. F. (2015). Models and neural bases of the believing process. *J. Behav. Brain Sci.* 5, 12–23. doi: 10.4236/jbbs.2015.51002

Teske, J. A. (2007). "Bindings of the will. The neuropsychology of subdoxastic faith", in *Humanity, World and God – Understanding and Actions*, eds W. B. Dress, H. Meisinger, and T. A. Smedes (Lund: Lund University), 27–44.

Thorn, C. A., and Graybiel, A. M. (2014). Differential entrainment and learning-related dynamics of spike and local field potential activity in the sensorimotor and associative striatum. *J. Neurosci.* 34, 2845–2859. doi: 10.1523/JNEUROSCI.1782-13.2014

Toga, A. W., and Mazziotta, J. C. (2000). *Brain Mapping. The Systems*. San Diego, CA: Acadamic Press.

Van Overwalle, F. (2009). Social cognition and the brain: a metaanalysis. *Hum. Brain Mapp.* 30, 829–858. doi: 10.1002/hbm.20547

Veit, R., Konicar, L., Klinzing, J. G., Barth, B., Yilmaz, Ö., and Birbaumer, N. (2013). Deficient fear conditioning in psychopathy as a function of interpersonal and affective disturbances. *Front. Hum. Neurosci.* 7:706. doi: 10.3389/fnhum.2013.00706

Wegner, D. M. (2003). The mind's best trick: how we experience conscious will. *Trends Cogn. Sci.* 7, 65–69. doi: 10.1016/S1364-6613(03)00002-0

Young, G. B., and Pigott, S. E. (1999). Neurobiological basis of consciousness. *Arch. Neurol.* 56, 153–157. doi: 10.1001/archneur.56.2.153

Conflict of Interest Statement: 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.

Copyright © 2017 Angel and Seitz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 July 2017
doi: 10.3389/fpsyg.2017.01253

Learning about Expectation Violation from Prediction Error Paradigms – A Meta-Analysis on Brain Processes Following a Prediction Error

Lisa D'Astolfo and Winfried Rief*

Department of Clinical Psychology and Psychotherapy, Philipps University of Marburg, Marburg, Germany

Modifying patients' expectations by exposing them to expectation violation situations (thus maximizing the difference between the expected and the actual situational outcome) is proposed to be a crucial mechanism for therapeutic success for a variety of different mental disorders. However, clinical observations suggest that patients often maintain their expectations regardless of experiences contradicting their expectations. It remains unclear which information processing mechanisms lead to modification or persistence of patients' expectations. Insight in the processing could be provided by Neuroimaging studies investigating prediction error (PE, i.e., neuronal reactions to non-expected stimuli). Two methods are often used to investigate the PE: (1) paradigms, in which participants passively observe PEs ("passive" paradigms) and (2) paradigms, which encourage a behavioral adaptation following a PE ("active" paradigms). These paradigms are similar to the methods used to induce expectation violations in clinical settings: (1) the confrontation with an expectation violation situation and (2) an enhanced confrontation in which the patient actively challenges his expectation. We used this similarity to gain insight in the different neuronal processing of the two PE paradigms. We performed a meta-analysis contrasting neuronal activity of PE paradigms encouraging a behavioral adaptation following a PE and paradigms enforcing passiveness following a PE. We found more neuronal activity in the striatum, the insula and the fusiform gyrus in studies encouraging behavioral adaptation following a PE. Due to the involvement of reward assessment and avoidance learning associated with the striatum and the insula we propose that the deliberate execution of action alternatives following a PE is associated with the integration of new information into previously existing expectations, therefore leading to an expectation change. While further research is needed to directly assess expectations of participants, this study provides new insights into the information processing mechanisms following an expectation violation.

Keywords: expectation violation, prediction error, fMRI, meta-analysis, striatum, insula

OPEN ACCESS

Edited by:
Karin Meissner, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Stephan Geuter, University of Colorado Boulder, United States
Florian Beissner, Hannover Medical School, Germany

****Correspondence:***
Lisa D'Astolfo
lisa.dastolfo@uni-marburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 October 2016*
Accepted: *10 July 2017*
Published: *28 July 2017*

Citation:
D'Astolfo L and Rief W (2017) Learning about Expectation Violation from Prediction Error Paradigms – A Meta-Analysis on Brain Processes Following a Prediction Error. Front. Psychol. 8:1253. doi: 10.3389/fpsyg.2017.01253

INTRODUCTION

Patients' expectations have a great influence on their treatment and outcomes in psychotherapy (Greenberg et al., 2006), medical conditions as well as in patients undergoing surgery (Auer et al., 2016; Rief and Glombiewski, 2016). In addition, negative expectations about psychological interventions may lead to negative effects of psychotherapy (Ladwig et al., 2014). Rief et al. (2015) have proposed to consider dysfunctional expectations to be core features of mental disorders. It has been argued that dysfunctional behavior is guided by dysfunctional expectations of situational associations and outcomes. Hence, behavioral therapy would only be successful if there is a change of the dysfunctional expectations guiding the behavior. These dysfunctional expectations are pre-existing assumptions about contingencies with a high subjective associative strength, i.e., subjective certainty. Patients would have to experience an expectation violation, i.e., a state, in which the expected outcome and the actual outcome differ, to induce a change in their expectations about the contingencies. This corresponds to a relearning of the contingencies, i.e., a state, in which they perceive a difference between expected outcome and the actual outcome, which would induce a change in their expectations about the contingencies. It is hypothesized that depending on various information processing variables, expectations might either be changed or maintained after an expectation violation situation. Thus, the relearning is either successful and persists over time or the relearning might be only temporary or depending on contextual factors.

The particular mechanisms underlying the information processing and the persistence and change of expectations have remained unclear. Clinical observations suggests that patients with mental disorders are particularly resistant to expectation change and the perception on expectation violations (Rief et al., 2015; Rief and Glombiewski, 2016). There are promising new approaches examining immunization as one of the processing strategies following expectation violation (Kube et al., 2016). This could explain why even after a successful expectation violation, the expectation is not changed. The patients perceive the violation of their pre-existing expectation but attribute the situation to contextual factors, e.g., the setting. Thus, the confrontation with an aversive stimulus with aim of reducing an emotional response, as is commonly used in psychotherapeutic settings, might not always be enough to induce a persistent expectation change. Craske et al. (2014) proposed methods of maximizing such exposure techniques, which are supposed to increase the inhibitory learning of the old expectation about the contingencies. One of these methods is the active testing of the pre-existing expectation. This is suggested to facilitate the relearning of the contingencies and to stabilize the newly learned expectation, thus inducing an expectation change.

The change of dysfunctional expectations is theorized as a crucial mechanism for therapeutic success. The investigation of cognitive processes facilitating an expectation change vs. maintenance following an expectation violation might pose a promising approach for cognitive behavioral therapy. Thus, we propose to compare the cognitive processing of a more passive confrontation with the aim of reducing an emotional response and an active approach by testing the expectation.

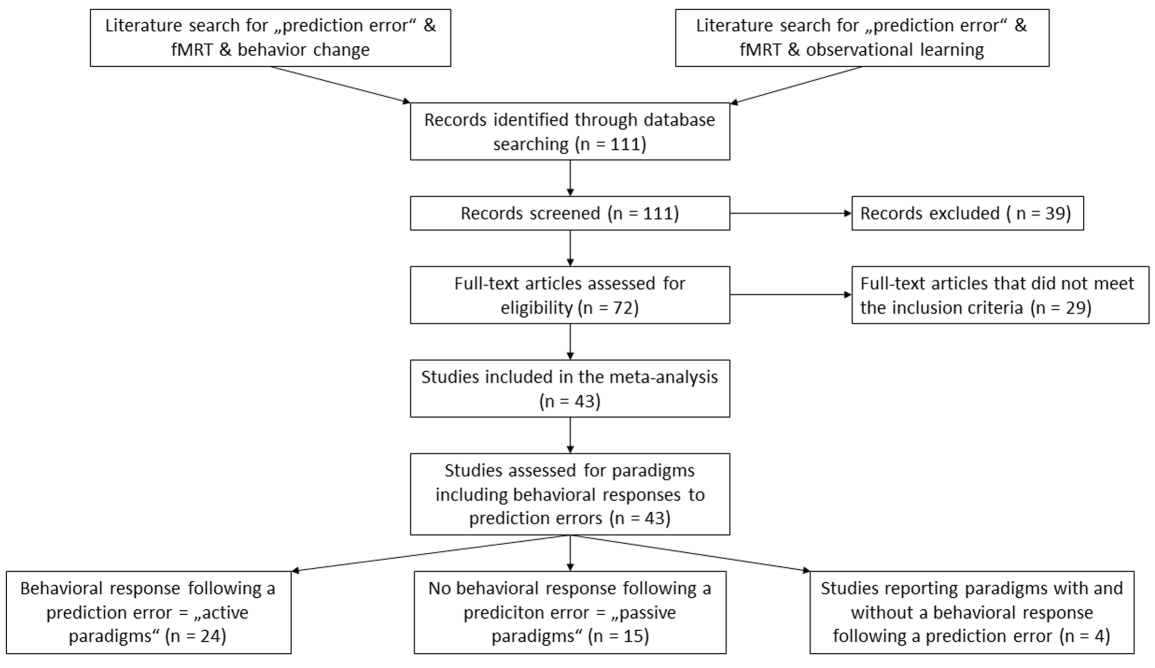


FIGURE 1 | Overview of the literature selection process.

TABLE 1 | Overview of the prediction error studies included in the meta-analysis.

Reference	Number of subjects	Task	Behavioral adaptation possible?
Abler et al., 2006	11	Reward probability	Yes
Behrens et al., 2007	18	Reward probability	Yes
Cohen, 2007	17	Reward probability	Yes and No
Daniel and Pollmann, 2012	18	Visual classification	Yes and No
Delgado et al., 2008	11	Aversive conditioning	No
den Ouden et al., 2010	20	Auditory classification	Yes
Diuk et al., 2013	28	Reward probability	Yes
Gläscher et al., 2010	18	Markov decision task	Yes
Gläscher et al., 2009	20	Reversal learning	Yes
Gradin et al., 2011	20	Reward learning	Yes
Ham et al., 2013	35	Simon Task	Yes
Hampton et al., 2006	16	Reversal learning	Yes
Haruno and Kawato, 2006	20	Association learning	Yes
Hester et al., 2010	16	Association learning	Yes
Hester et al., 2008	17	Spatial learning task	Yes and No
Kahnt et al., 2011	20	Orientation discrimination	Yes
Kim et al., 2006	16	Instrumental choice	Yes
Klein et al., 2007	26	Probabilistic learning	Yes
Kumar et al., 2008	18	Reward conditioning	No
Landmann et al., 2007	16	Trial-and-error learning	Yes
Li et al., 2011	17	Fear conditioning	No
Li et al., 2006	46	Reward probability	Yes
Limongi et al., 2013	15	Michotte's Launching effect	No
McClure et al., 2003	18	Reward conditioning	No
Metereau and Dreher, 2013	20	Pavlovian conditioning	No
Morris et al., 2012	16	Reward conditioning	Yes
Murray et al., 2007	12	Reward learning	Yes
Niv et al., 2012	16	Reward probability	Yes
O'Doherty et al., 2004	12	Reward conditioning	Yes
Ploghaus et al., 2000	12	Fear conditioning	No
Ramnani et al., 2004	6	Instrumental conditioning	No
Rodriguez et al., 2006	15	Visual classification	Yes
Schiller et al., 2008	17	Fear conditioning	No
Schlerf et al., 2012	10	Motoric learning	No
Schonberg et al., 2010	17	Reward probability	Yes
Seymour et al., 2007	24	Reward conditioning	No
Seymour et al., 2005	19	Aversive conditioning	No
Spoormaker et al., 2011	40	Fear conditioning	No
Takemura et al., 2011	23	Reward conditioning	No
Tobler et al., 2006	22	Reward blocking	Yes and No
Valentin and O'Doherty, 2009	17	Reward probability	Yes and No
Watanabe et al., 2013	20	Reward probability	Yes

The neuroimaging research on learning provides experimentally designed expectation violations. One of the concepts consistently associated with successful learning is the so-called prediction error, i.e., the neurological response to an unexpected stimulus. Learning research has mainly focused on reinforcement learning, whereby the expectations comprises predictions about reward and/or punishment (Karuza et al., 2014). Many studies use partial reinforcement or probabilistic learning paradigms. It can be argued that changes in behavioral strategies in these paradigms also reflect changes in underlying expectations regarding the contingencies of reward and punishment. Hence, in paradigms, in which no behavioral adaptation is necessary, i.e., a passive observation of contingencies, might diminish the attention on expectation violations. We argue that participants in both paradigms compute prediction errors and their relearning of the contingencies is successful. In alignment with the approach by Craske et al. (2014) to maximize inhibitory learning by actively testing the expectation, we hypothesize a different cognitive processing of "active" paradigms, which encourage a behavioral adaptation and

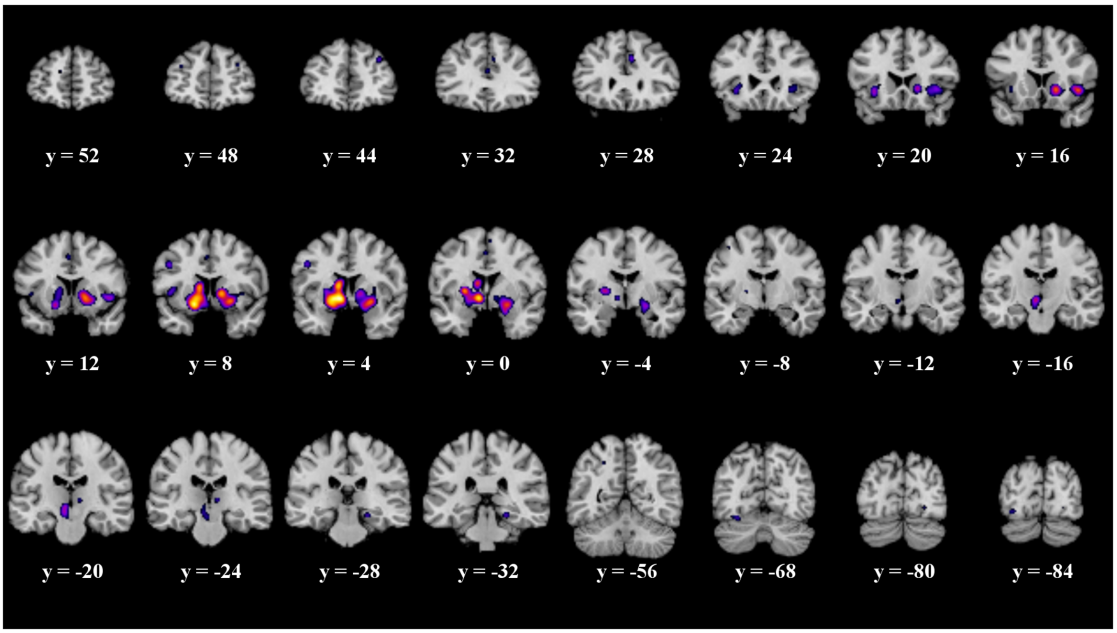


FIGURE 2 | Overview of the meta-analysis results for all studies ($p < 0.05$, FDR). The significant clusters are comprised of activity in the superior frontal gyrus, middle frontal gyrus, anterior cingulate, cingulate gyrus, claustrum, insula, caudate head, precentral gyrus, putamen, lateral globus pallidus, caudate body, red nucleus, thalamus, parahippocampal gyrus, superior parietal lobule, declive, lingual gyrus and the fusiform gyrus (from left to right). MNI coordinates are presented below each coronal view.

"passive" paradigms, in which contingencies are observed. Since the concepts of prediction error and expectation violation are identical in matters of meaning for the preexisting expectation, it seems likely that clinical research can benefit from an insight of neuroimaging research on prediction error. Examining the functional magnetic resonance imaging (fMRI) results provided by research on prediction error might provide insights in the cognitive processes associated with the information processing during expectation violations.

Our aim is to review fMRI studies investigating two different prediction error paradigms. The first paradigm encourages strategic behavioral changes throughout the course of experiments while the second one requires a passive observation. A contrast analysis will be performed to identify differences in brain activity between these two paradigm categories. A meta-analysis summarized the current findings on brain areas associated with prediction error (Garrison et al., 2013). They found a consistent association of the pallidum, the striatum and medio-frontal structures with prediction error. These structures are also associated with the fronto-striatal circuits. The circuit is defined as circular connections between the caudate nucleus, putamen, thalamus and prefrontal regions (Leh et al., 2007). Dysfunctions in this circuit are associated with impaired behavioral adaptation such as poor set shifting performance, e.g., in a go/no-go tests or stimulus-bound behavior (Mega and Cummings, 1994). Several disorders are linked to fronto-striatal circuit dysfunctions, such as Huntington's disease (Beste et al., 2012), Parkinson's Disease (Owen, 2004) and obsessive-compulsive disorder (Maltby et al., 2005; Marsh et al., 2014). All clinical pictures are associated with behavioral and cognitive perseverations (Mega and Cummings, 1994). It therefore seems likely to assume the fronto-striatal circuit to be involved in the expectation violation processing and the resulting expectation and behavioral adaptation. We will perform a meta-analysis involving prediction error followed by a behavioral adaptation to an uncertain environment. We expect a consistent activation in the striatum and media-frontal areas.

MATERIALS AND METHODS

Literature Selection

We conducted a systematic literature search to identify neuroimaging studies of prediction error using PubMed[1], Web of Science[2], and Neurosynth[3] databases. We searched for articles in the English language using the keywords "prediction error" AND "fMRI" and did not specify a time span for date of publication. The search revealed 8'610 results as of July 2016. To narrow the results, a second search was performed using the keywords "prediction error" AND "fMRI" AND "behavior change" as well as "prediction error" AND "fMRI" AND "observational learning". Again, no time span was specified. These searches revealed 111 results and four results, respectively. The abstracts of these articles were examined to select potential matches for our inclusion criteria. We also scanned the reference lists of

[1]http://www.ncbi.nlm.nih.gov/pubmed
[2]http://apps.webofknowledge.com/
[3]http://neurosynth.org/

TABLE 2 | Details of the clusters revealed by the analysis across all studies.

Cluster		MNI coordinates			Cluster size [mm^3]
		X	*Y*	*Z*	
Caudate Head	R	12	8	0	10'712
	R	16	14	−4	
Putamen	R	24	6	−8	
	R	28	4	6	
Insula	R	42	16	−4	
Lateral Globus Pallidus	R	22	0	−12	
Claustrum	R	30	22	−4	
Medial Globus Pallidus	L	−10	2	−4	9'296
Lateral Globus Pallidus	L	−16	6	−6	
Putamen	L	−18	6	−10	
	L	−26	0	4	
Caudate Body	L	−12	6	8	
Red Nucleus	L	−8	−18	−6	1'704
Claustrum	L	−32	22	−6	928
Precentral Gyrus	L	−44	8	4	528
	L	−46	6	34	488
Cingulate Gyrus	R	6	28	32	368
Parahippocampal Gyrus	R	22	−30	−14	312
Middle Frontal Gyrus	R	34	44	30	312
Superior Frontal Gyrus	R	32	48	24	
Declive	L	−30	−68	−18	272
Fusiform Gyrus	L	−34	−86	−10	232
Cingulate Gyrus	L	−4	10	42	232
Thalamus	R	8	−22	2	224
Anterior Cingulate	L	0	34	18	216
Cingulate Gyrus	L	−2	0	48	120
Superior Frontal Gyrus	L	−12	54	18	104
Lingual Gyrus	R	24	−82	−6	88
Precentral Gyrus	L	−40	−10	54	72
Medial Frontal Gyrus	R	4	0	60	64
Middle Frontal Gyrus	L	−32	48	24	56
Superior Parietal Lobule	L	−26	−56	44	56

Threshold Method = FDR; Thresholding Value = 0.05; Chosen min. cluster size = 50 mm^3; R, Right; L, Left.

the results to search for additional articles, which met our inclusion criteria. We retrieved the full text of 72 articles for further examination. We predefined study selection criteria to minimize ambiguousness in the study selection. The criteria can be requested of the corresponding author. Studies were included when they met the following criteria: (1) experimental prediction error paradigm and (2) report of voxel-wise-brain analysis for a prediction error main effect, which yielded a total of 59 articles. We excluded studies which did not report prediction error for healthy adults or used medication in their experiment ($n = 6$ studies excluded). We did this to include only prediction errors which arise from an unexpected change in contingencies in alignment with the clinical model. Of these studies, we precluded those articles failing to experimentally induce a prediction error by changing the contingencies between stimuli and outcome ($n = 10$ studies excluded). A flowchart of the selection process is shown in **Figure 1**. The studies included in the meta-analysis are listed in **Table 1**.

Contrast Selection

We included all analyses which contrasted prediction error brain activity with brain activity during expectation confirming trails or paradigm specific variations of these contrasts. Of the 43 studied that met all inclusion criteria, we included 60 contrasts in the analysis. If the coordinates were reported in Talairach space they were transformed to Montreal Neurologic Institute (MNI) space using the GingerAle software (Eickhoff et al., 2009, 2012), which utilizes the icbm2tal transform algorithm (Lancaster et al., 2007). In total, we included 446 foci into the analysis.

Activation Likelihood Estimation (ALE)

We performed an activation likelihood estimation (ALE) analysis using the Software GingerAle (Eickhoff et al., 2009, 2012). The algorithm assesses above-chance clustering between experiments, using a probability distribution centered at each of the foci used in the analysis. Since the spatial relationship is assumed to be fixed in each experiment, the ALE analysis infers random effects (Eickhoff et al., 2009). We used the algorithm described in Turkeltaub et al. (2012), which organizes the foci by subject group (as opposed to study affiliation). This prevents an influence of multiple foci from one experiment on the Meta-Analysis results (Turkeltaub et al., 2012). We performed three Meta-Analyses: (1) studies which encourage a behavioral strategic adaptation following a prediction error, (2) studies, which employed a passive observational paradigm, and (3) an analysis of all studies, which was necessary to perform the contrast analysis. In line with previous studies (Garrison et al., 2013), we defined a false discovery rate (FDR) method with $p < 0.05$ and a minimal cluster volume of 50 mm^3. We then performed a contrast analysis of the "active" behavioral subset and the "passive" observational study subset. This analysis allows the subtraction of two datasets to compare differences in brain activity between these two. To this end, a pooled dataset is created, which then serves as basis for two randomly created datasets with the same number of foci as the original datasets. A permutation of subtractions of simulated datasets are compared to the results of the original datasets. We used an uncorrected p-value $p < 0.05$ since the single analyses were already corrected with FDR (Eickhoff et al., 2011). We chose a minimal cluster volume of 50 mm^3 for the contrast analysis. Papaya[4] was used to superimpose the ALE cluster results on a T1 brain template (Colin27_T1_seg_MNI.nii[5]).

RESULTS

Meta-Analysis across All Studies of Prediction Error

Twenty-one significant clusters were identified by the ALE meta-analysis of all 43 studies. The results show activation in the right basal ganglia and the right insula (see details in **Figure 2** and **Table 2**). There was no clear indication of laterality in the main analysis.

[4]http://ric.uthscsa.edu/mango/papaya.html
[5]http://www.brainmap.org./ale/

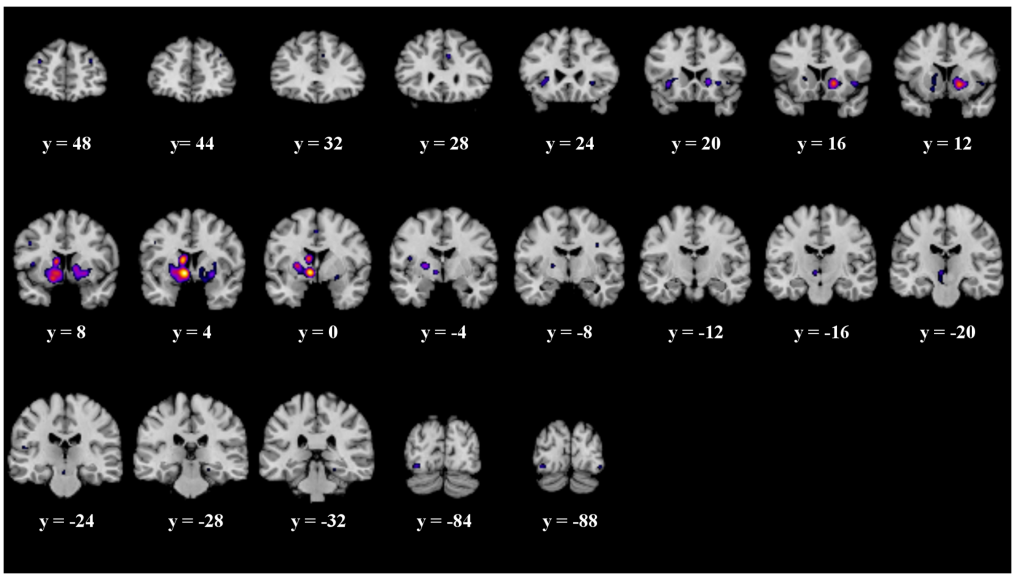


FIGURE 3 | Overview of the meta-analysis results for the studies reporting active behavioral paradigms ($p < 0.05$, FDR). The significant clusters are comprised of activity in the superior frontal gyrus, middle frontal gyrus, anterior cingulate, cingulate gyrus, claustrum, insula, caudate head, inferior frontal gyrus, caudate body, putamen, medial globus pallidus, thalamus, substancia nigra, red nucleus, transverse temporal gyrus, parahippocampal gyrus, fusiform gyrus and inferior occipital gyrus (from left to right). MNI coordinates are presented below each coronal view.

Meta-Analyses for Behavioral and Observational Paradigms

When analyzing all prediction error studies, which employed a behavioral reaction following a prediction error, the ALE meta-analysis revealed 17 significant clusters. We found activation in the striatum, the insula and the claustrum (see details in **Figure 3** and **Table 3**).

The ALE meta-analysis of all prediction error studies, which employed a passive paradigm revealed four significant clusters. We found activation in the putamen, the lateral globus pallidus, declive and the lingual gyrus (see details in **Figure 4** and **Table 3**).

In both analyses, no clear indication of laterality was found.

Subtraction Analysis

The details of the ALE subtraction analysis are shown in **Table 4** and **Figure 5**. In the contrast behavior – passive, we found five significant clusters, comprising parts of the striatum, the insula and the fusiform gyrus. There was a tendency of left sided structures to be more active in prediction error paradigms encouraging behavioral adaptation. We found no significant clusters in the contrast passive – behavior.

It is often suggested to apply corrected thresholds to the contrast analyses, such as a FDR threshold. Therefore, we replicated the subtraction analyses with more conservative thresholds. We applied a corrected FDR threshold of $p < 0.05$ to the subtraction analysis. We found no significant clusters in the contrast passive – behavior. The significant clusters of the contrast behavior – passive do not survive the corrected threshold.

DISCUSSION

We performed a subtraction analysis of two different prediction error paradigms. One encourages a behavioral adaptation to changing contingencies while the second paradigm requires a passive observation of contingencies. Our aim was to gain a better understanding of why and how psychological interventions focusing on expectation violation lead to behavioral changes in some but not all cases. Therefore, we analyzed differences in prediction error involving on one hand the execution of an action alternative and on the other hand no behavioral change. We wanted to identify cognitive processes being involved in underlying expectations about contingencies guiding the behavior. As a major result when contrasting studies employing the two paradigms discussed earlier, we found significantly more activation in the left medial globus pallidus, the left caudate body, the right caudate head and putamen as well as the left fusiform gyrus and the left insula.

All Studies of Prediction Error

When performing a meta-analysis containing all prediction error studies our results are in line with previous research (Garrison et al., 2013). We found activation in the striatum, the insula, thalamus as well as fronto-medial structures. The Putamen and the Caudate body are part of the striatum whose association with memory processes is consistent with previous literature (Grahn et al., 2009; Provost et al., 2015). The insula has been associated primarily with fear conditioning (Kircher et al., 2013) but also with reinforcement learning for reward (Lawrence et al., 2014) as well for avoidance learning (Palminteri et al., 2012).

TABLE 3 | Details of the clusters revealed by the analyses of the behavioral and passive studies.

Cluster		MNI coordinates			Cluster size [mm^3]
		X	*Y*	*Z*	
Behavioral paradigms					
Medial Globus Pallidus	L	−10	2	−4	8'528
Caudate Body	L	−10	4	12	
Putamen	L	−16	0	4	
	L	−20	16	2	
Claustrum	L	−32	22	−4	
Caudate Head	R	16	14	−4	4'896
Claustrum	R	30	22	−4	
Putamen	R	28	4	6	
Thalamus	L	8	−18	−4	608
Substantia Nigra	L	−8	−20	−14	
Red Nucleus	L	−4	−22	−18	
Insula	R	42	16	−4	480
Fusiform Gyrus	L	−34	−86	−10	344
Cingulate Gyrus	R	6	28	30	208
Insula	L	−32	48	24	144
Inferior Frontal Gyrus	L	−46	6	32	136
Insula	L	−42	−4	14	120
Parahippocampal Gyrus	R	22	−30	−14	104
Cingulate Gyrus	L	−2	0	48	104
Superior Frontal Gyrus	R	32	48	24	96
Inferior Occipital Gyrus	R	38	−88	−12	88
Transverse temporal Gyrus	L	−52	−24	12	64
Anterior Cingulate	L	−2	34	16	64
Middle Frontal Gyrus	R	36	46	32	56
Passive paradigms					
Putamen	L	−20	6	10	720
Lateral Globus Pallidus	R	20	−2	−12	680
Putamen	R	22	6	−10	
Declive	L	−30	−68	−18	80
Lingual Gyrus	R	24	−82	−6	80

Threshold Method = FDR; Thresholding Value = 0.05; Chosen min. cluster size = 50 mm^3; R, Right; L, Left.

Consistent with our hypothesis we also found activation in the areas associated with the fronto-striatal circuits (Leh et al., 2007). In addition to the striatum, we found activation in the globus pallidus, the thalamus and frontal structures, i.e., the left superior, media and middle frontal gyrus.

Prediction Error Followed by Behavioral Adaptation or Passive Observation

When contrasting the differences in neuronal activity of prediction errors computed in active behavioral adaptation and passive observational paradigms we found higher activation in the striatum, the insula and the fusiform gyrus.

The medial globus pallidus is part of the four corticostriatal loops, which are responsible for executive function, visual processing, motor function and motivational evaluation (Seger, 2006). It serves as an output nucleus of the basal ganglia and projects to the thalamus, the centromedian nucleus, and the pedunculopontine nucleus (Nauta and Mehler, 1966). These structures are associated with goal-directed motor actions as well as reward learning and evaluation (Hong and Hikosaka, 2008; Haber and Knutson, 2009; Sescousse et al., 2013).

The putamen is associated with novel motoric executions as well as in ambiguous action tendencies, i.e., if the best motoric strategy is unclear (Grahn et al., 2009). Moreover, due to findings of strong connectivity of the putamen with prefrontal regions, it is suggested that the putamen has a cognitive rather than solely motoric function (Provost et al., 2015).

The caudate body has been shown to be involved in cognitive tasks such as categorization and reward information assessment in monkeys (Yanike and Ferrera, 2014) as well as in humans (Packard and Knowlton, 2002). Further, it has been suggested, that the caudate nucleus is involved in evaluating outcomes post-decision (Badre, 2012; Kepecs and Mainen, 2012).

Most studies do not specifically differentiate between distinct parts of the striatum, but investigate the striatum in its entirety. The striatum has been associated with strategizing in avoidance learning (Palminteri et al., 2012), failure or success to learn associations in instrumental conditioning (Schönberg et al., 2007; Horga et al., 2015), decision making and motor initiation (Nagano-Saito et al., 2014).

The insula is associated with the perception and processing of interoception of emotional states (Zaki et al., 2012; Simmons et al., 2013).

The fusiform gyrus is associated with facial and body recognition (Peelen and Downing, 2005) as well as a sensitivity to visual words (McCandliss et al., 2003). The area of the fusiform gyrus showing peak activation is also associated with object recognition (Bar et al., 2001).

The functions of these areas can be incorporated into the processing of prediction errors computed in a behavioral paradigm. The higher activation of the putamen in prediction errors with behavioral changes might be due to the determination of a novel motoric behavior and its initiation. Due to its' evaluative properties, the caudate nucleus could function as a constant evaluation unit, comparing expected and actual outcomes. The involvement of the insula cannot be explained by the emotional valence of the stimuli used in the studies, since not all the studies comprising the insula cluster contained emotional content, such as negative feedback. They share, however, a high level of uncertainty in their paradigms, e.g., temporal uncertainty or ambiguous stimuli or categories. The processing of uncertainty has also been shown to be associated with insula activity (Simmons et al., 2008; Sarinopoulos et al., 2010), which could be interpreted as an aversive and thus emotional state.

Integration into a Clinical Model of Expectation Change and Persistence

Rief et al.'s (2015) model proposes that following an expectation violation, various information processing mechanisms decide whether an expectation is changed and integrated or maintained and reinforced. In order to shed light on the cognitive processes involved in an expectation change following an expectation violation, we investigated the brain areas more active in

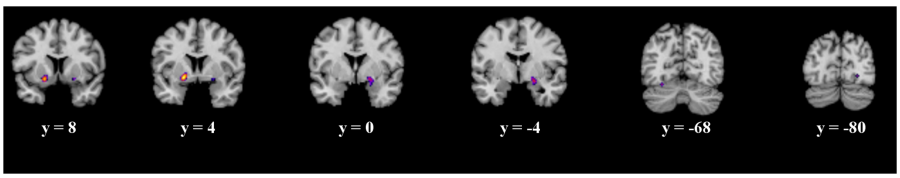


FIGURE 4 | Overview of the meta-analysis results for the studies reporting passive observational paradigms ($p < 0.05$, FDR). The significant clusters are comprised of activity in the putamen, lateral globus pallidus, declive and the lingual gyrus (from left to right). MNI coordinates are presented below each coronal view.

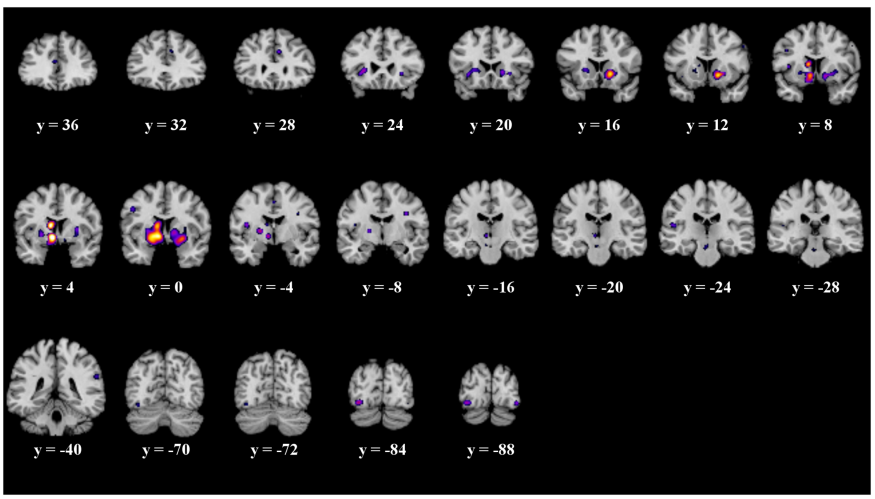


FIGURE 5 | Overview of the meta-analysis results of the behavior – no behavior subtraction analysis ($p < 0.05$, uncorrected). The significant clusters are comprised of activity in the putamen, caudate head, caudate body, insula, medial globus pallidus, and the fusiform gyrus (from left to right). MNI coordinates are presented below each coronal view.

TABLE 4 | Details of the clusters revealed by subtraction analysis.

Cluster		MNI coordinates			Cluster size [mm^3]
		X	*Y*	*Z*	
Behavior – no behavior					
Medial Globus Pallidus	L	−8	2	−12	1'296
Caudate Body	L	−6	2	18	848
	L	−12	4	18	
	L	−10	8	16	
	L	−10	2	14	
	L	−6	4	12	
	L	−14	−2	16	
Caudate Head	R	12	12	−8	704
Putamen	R	20	16	−4	
	R	24	16	−2	
Fusiform Gyrus	L	−38	−84	−12	200
	L	−34	−84	−8	
Insula	L	−42	−6	16	72
No behavior – behavior					
n.s.					

Threshold Method = Uncorrected P-value; Thresholding Value = 0.05; Thresholding Permutations = 10000; Chosen min. cluster size = 50 mm^3; R, Right; L, Left.

paradigms encouraging a behavioral strategic change following a prediction error.

The striatum might be involved in learning the specific contingencies between stimulus and outcome. This might eventually form an expectation about the action strategies resulting in a rewarding outcome. However, when facing an expectation violation, the caudate body might signal a non-rewarding outcome, even though the same behavioral strategy has been employed. On the other hand, if the environment encourages a passive behavior, i.e., no action has to be taken following an expectation violation, the individual is not required to determine a behavioral alternative. The difference in expectation and outcome could be solved by mechanisms such as immunization, leading to an expectation persistence (Kube et al., 2016). In contrast, if the environment encourages or even enforces the use of action alternatives, e.g., an active prediction error paradigm or a therapeutic setting, a behavioral reaction to the situation would be necessary. In such a case, the putamen could be involved in determining a novel behavior and initiate the action alternative by projecting to the medial globus pallidus. This structure could then initiate the motoric aspect of the action alternative. The thalamus, the centromedian nucleus, and the pedunculopontine nucleus could be involved in assessing the reward when employing the new behavior. If the

action alternative leads to a satisfying result, i.e., a rewarding outcome, the behavior is integrated and leads to an expectation change.

The involvement of the insula especially in prediction error paradigms encouraging behavioral adaptation suggests an emotional component to be important. Due to its association with avoidance learning, the insula might be involved in assessing aversive outcomes following an action alternative. This contrasts with the reward assessment of thalamus, centromedian nucleus and pedunculopontine nucleus. It might be possible, that the avoidance of an unwanted outcome, e.g., a negative emotional state, is as important as the gain of a rewarding outcome. The aversion of a negative emotional state might be a rewarding outcome in itself which has to be considered when assessing the reward of an action alternative.

Limitations

A few methodological limitations have to be considered. First, the studies we included used various paradigms, showing a wide range of stimuli, tasks and underlying mathematical models. However, there is evidence of different brain activity involved in various types of learning and in particular in model-based (i.e., goal-directed actions) vs. model-free (i.e., habit-based actions) approaches (Maia, 2009; Wunderlich et al., 2012). Moreover, the ALE meta-analysis itself has a few limitations. Coordinate-based analyses accumulate power across studies (Costafreda, 2009) and cannot reproduce the same quality in results as image-based meta-analyses (Salimi-Khorshidi et al., 2009). A third limitation is that the results of the subtraction analysis do not survive a FDR corrected threshold. Considering this restriction, the results of the contrast analysis have to be interpreted with caution. Eickhoff et al. (2016) recommend a minimal sample size of 17 studies for the ALE meta-analysis. We could only include 19 studies using a passive observational paradigm. This suggests that the statistical power may be rather small for the subtraction analysis, explaining why our results did not survive the FDR corrected threshold. For future research, it is necessary to repeat the analyses with a larger study sample to increase the statistical power. This will allow a more decisive analysis of the differences in neurological activity between active behavioral and passive observational prediction error paradigms.

CONCLUSION

This meta-analysis sheds light into the cognitive processes involved in the execution of action alternatives following an expectation violation. The information processing involved is strongly associated with reward evaluation of newly found behavioral adaptations. However, further research is needed in order to explicitly investigate the expectations of participants of prediction error paradigms regarding their behavioral strategies.

AUTHOR CONTRIBUTIONS

LD: Did the major part of the work regarding conception and methodology of the article; performed the analyses and the evaluation of the results; approves of the manuscript being published; agrees on being accountable for all aspects of the work, ensuring that questions regarding quality and accuracy of the work are investigated appropriately and resolved. WR: Substantially contributed to the conception of the article, revised the manuscript critically and contributed to the content; approves of the manuscript being published; agrees on being accountable for all aspects of the work, ensuring that questions regarding quality and accuracy of the work are investigated appropriately and resolved.

ACKNOWLEDGMENT

The authors thank Tobias Kube for his constructive correction of the manuscript.

REFERENCES

Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage* 31, 790–795. doi: 10.1016/j.neuroimage.2006.01.001

Auer, C. J., Glombiewski, J. A., Doering, B. K., Winkler, A., Laferton, J. A. C., Broadbent, E., et al. (2016). Patients' expectations predict surgery outcomes: a meta-analysis. *Int. J. Behav. Med.* 23, 49–62. doi: 10.1007/s12529-015-9500-4

Badre, D. (2012). Opening the gate to working memory. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19878–19879. doi: 10.1073/pnas.1216902109

Bar, M., Tootell, R. B. H., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., et al. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuron* 29, 529–535. doi: 10.1016/S0896-6273(01)00224-0

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221. doi: 10.1038/nn1954

Beste, C., Ness, V., Lukas, C., Hoffmann, R., Stüwe, S., Falkenstein, M., et al. (2012). Mechanisms mediating parallel action monitoring in fronto-striatal circuits. *Neuroimage* 62, 137–146. doi: 10.1016/j.neuroimage.2012.05.019

Cohen, M. X. (2007). Individual differences and the neural representations of reward expectation and reward prediction error. *Soc. Cogn. Affect. Neurosci.* 2, 20–30. doi: 10.1093/scan/nsl021

Costafreda, S. (2009). Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies. *Front. Neuroinformat.* 3:33. doi: 10.3389/neuro.11.033.2009

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., and Vervliet, B. (2014). Maximizing exposure therapy: an inhibitory learning approach. *Behav. Res. Ther.* 58, 10–23. doi: 10.1016/j.brat.2014.04.006

Daniel, R., and Pollmann, S. (2012). Striatal activations signal prediction errors on confidence in the absence of external feedback. *Neuroimage* 59, 3457–3467. doi: 10.1016/j.neuroimage.2011.11.058

Delgado, M. R., Li, J., Schiller, D., and Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philos. Trans. R. Soc. B: Biol. Sci.* 363, 3787–3800. doi: 10.1098/rstb.2008.0161

den Ouden, H. E. M., den Daunizeau, J., Roiser, J., Friston, K. J., and Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *J. Neurosci.* 30, 3210–3219. doi: 10.1523/JNEUROSCI.4458-09.2010

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., and Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *J. Neurosci.* 33, 5797–5805. doi: 10.1523/JNEUROSCI.5445-12.2013

Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., and Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *Neuroimage* 59, 2349–2361. doi: 10.1016/j.neuroimage.2011.09.017

Eickhoff, S. B., Bzdok, D., Laird, A. R., Roski, C., Caspers, S., Zilles, K., et al. (2011). Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *Neuroimage* 57, 938–949. doi: 10.1016/j.neuroimage.2011.05.021

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., and Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* 30, 2907–2926. doi: 10.1002/hbm.20718

Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., et al. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage* 137, 70–85. doi: 10.1016/j.neuroimage.2016.04.072

Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* 37, 1297–1310. doi: 10.1016/j.neubiorev.2013.03.023

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595. doi: 10.1016/j.neuron.2010.04.016

Gläscher, J., Hampton, A. N., and O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cereb. Cortex* 19, 483–495. doi: 10.1093/cercor/bhn098

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., et al. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain* 134, 1751–1764. doi: 10.1093/brain/awr059

Grahn, J. A., Parkinson, J. A., and Owen, A. M. (2009). The role of the basal ganglia in learning and memory: neuropsychological studies. *Behav. Brain Res.* 199, 53–60. doi: 10.1016/j.bbr.2008.11.020

Greenberg, R. P., Constantino, M. J., and Bruce, N. (2006). Are patient expectations still relevant for psychotherapy process and outcome? *Clin. Psychol. Rev.* 26, 657–678. doi: 10.1016/j.cpr.2005.03.002

Haber, S. N., and Knutson, B. (2009). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* 35, 4–26. doi: 10.1038/npp.2009.129

Ham, T., Leff, A., Boissezon, X., de Joffe, A., and Sharp, D. J. (2013). Cognitive control and the salience network: an investigation of error processing and effective connectivity. *J. Neurosci.* 33, 7091–7098. doi: 10.1523/JNEUROSCI.4692-12.2013

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* 26, 8360–8367. doi: 10.1523/JNEUROSCI.1010-06.2006

Haruno, M., and Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *J. Neurophysiol.* 95, 948–959. doi: 10.1152/jn.00382.2005

Hester, R., Barre, N., Murphy, K., Silk, T. J., and Mattingley, J. B. (2008). Human medial frontal cortex activity predicts learning from errors. *Cereb. Cortex* 18, 1933–1940.

Hester, R., Murphy, K., Brown, F. L., and Skilleter, A. J. (2010). Punishing an error improves learning: the influence of punishment magnitude on error-related neural activity and subsequent learning. *J. Neurosci.* 30, 15600–15607. doi: 10.1523/JNEUROSCI.2565-10.2010

Hong, S., and Hikosaka, O. (2008). The globus pallidus sends reward-related signals to the lateral habenula. *Neuron* 60, 720–729. doi: 10.1016/j.neuron.2008.09.035

Horga, G., Maia, T. V., Marsh, R., Hao, X., Xu, D., Duan, Y., et al. (2015). Changes in corticostriatal connectivity during reinforcement learning in humans. *Hum. Brain Mapp.* 36, 793–803. doi: 10.1002/hbm.22665

Kahnt, T., Grueschow, M., Speck, O., and Haynes, J.-D. (2011). Perceptual learning and decision-making in human medial frontal cortex. *Neuron* 70, 549–559. doi: 10.1016/j.neuron.2011.02.054

Karuza, E. A., Emberson, L. L., and Aslin, R. N. (2014). Combining fMRI and behavioral measures to examine the process of human learning. *Neurobiol. Learn. Mem.* 109, 193–206. doi: 10.1016/j.nlm.2013.09.012

Kepecs, A., and Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. B* 367, 1322–1337. doi: 10.1098/rstb.2012.0037

Kim, H., Shimojo, S., and O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol.* 4:e233. doi: 10.1371/journal.pbio.0040233

Kircher, T., Arolt, V., Jansen, A., Pyka, M., Reinhardt, I., Kellermann, T., et al. (2013). Effect of cognitive-behavioral therapy on neural correlates of fear conditioning in panic disorder. *Biol. Psychiatry* 73, 93–101. doi: 10.1016/j.biopsych.2012.07.026

Klein, T. A., Neumann, J., Reuter, M., Hennig, J., von Cramon, D. Y., and Ullsperger, M. (2007). Genetically determined differences in learning from errors. *Science* 318, 1642–1645.

Kube, T., D'Astolfo, L., Glombiewski, J. A., Doering, B. K., and Rief, W. (2016). Focusing on situation-specific expectations in major depression as basis for behavioral experiments - development of the Depressive Expectations Scale (DES). *Psychol. Psychother.* doi: 10.1111/papt.12114 [Epub ahead of print].

Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., and Steele, J. D. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain* 131, 2084–2093. doi: 10.1093/brain/awn136

Ladwig, I., Rief, W., and Nestoriuc, Y. (2014). What are the risks and side effects of psychotherapy?-development of an inventory for the assessment of negative effects of psychotherapy (INEP). *Verhaltenstherapie* 24, 252–264.

Lancaster, J. L., Tordesillas-Gutiérrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., et al. (2007). Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum. Brain Mapp.* 28, 1194–1205. doi: 10.1002/hbm.20345

Landmann, C., Dehaene, S., Pappata, S., Jobert, A., Bottlaender, M., Roumenov, D., et al. (2007). Dynamics of prefrontal and cingulate activity during a reward-based logical deduction task. *Cereb. Cortex* 17, 749–759. doi: 10.1093/cercor/bhk028

Lawrence, E. J., Su, L., Barker, G. J., Medford, N., Dalton, J., Williams, S. C. R., et al. (2014). Self-regulation of the anterior insula: reinforcement learning using real-time fMRI neurofeedback. *Neuroimage* 88, 113–124. doi: 10.1016/j.neuroimage.2013.10.069

Leh, S. E., Ptito, A., Chakravarty, M. M., and Strafella, A. P. (2007). Fronto-striatal connections in the human brain: a probabilistic diffusion tractography study. *Neurosci. Lett.* 419, 113–118. doi: 10.1016/j.neulet.2007.04.049

Li, J., McClure, S. M., King-Casas, B., and Montague, P. R. (2006). Policy adjustment in a dynamic economic game. *PLoS ONE* 1:e103. doi: 10.1371/journal.pone.0000103

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., and Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nat. Neurosci.* 14, 1250–1252. doi: 10.1038/nn.2904

Limongi, R., Sutherland, S. C., Zhu, J., Young, M. E., and Habib, R. (2013). Temporal prediction errors modulate cingulate–insular coupling. *Neuroimage* 71, 147–157. doi: 10.1016/j.neuroimage.2012.12.078

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: successes and challenges. *Cogn. Affect. Behav. Neurosci.* 9, 343–364. doi: 10.3758/CABN.9.4.343

Maltby, N., Tolin, D. F., Worhunsky, P., O'Keefe, T. M., and Kiehl, K. A. (2005). Dysfunctional action monitoring hyperactivates frontal–striatal circuits in obsessive–compulsive disorder: an event-related fMRI study. *Neuroimage* 24, 495–503. doi: 10.1016/j.neuroimage.2004.08.041

Marsh, R., Horga, G., Parashar, N., Wang, Z., Peterson, B. S., and Simpson, H. B. (2014). Altered activation in fronto-striatal circuits during sequential processing of conflict in unmedicated adults with obsessive-compulsive disorder. *Biol. Psychiatry* 75, 615–622. doi: 10.1016/j.biopsych.2013.02.004

McCandliss, B. D., Cohen, L., and Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn. Sci.* 7, 293–299. doi: 10.1016/S1364-6613(03)00134-7

McClure, S. M., Berns, G. S., and Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38, 339–346. doi: 10.1016/S0896-6273(03)00154-5

Mega, M., and Cummings, J. (1994). Frontal-subcortical circuits and neuropsychiatric disorders. *J. Neuropsychiatry Clin. Neurosci.* 6, 358–370. doi: 10.1176/jnp.6.4.358

Metereau, E., and Dreher, J.-C. (2013). Cerebral correlates of salient prediction error for different rewards and punishments. *Cereb. Cortex* 23, 477–487. doi: 10.1093/cercor/bhs037

Morris, R. W., Vercammen, A., Lenroot, R., Moore, L., Langton, J. M., Short, B., et al. (2012). Disambiguating ventral striatum fMRI-related bold signal during reward prediction in schizophrenia. *Mol. Psychiatry* 17, 280–289. doi: 10.1038/mp.2011.75

Murray, G. K., Corlett, P. R., Clark, L., Pessiglione, M., Blackwell, A. D., Honey, G., et al. (2007). Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Mol. Psychiatry* 13, 267–276. doi: 10.1038/sj.mp.4002058

Nagano-Saito, A., Martinu, K., and Monchi, O. (2014). Function of basal ganglia in bridging cognitive and motor modules to perform an action. *Front. Neurosci.* 8:187. doi: 10.3389/fnins.2014.00187

Nauta, W. J. H., and Mehler, W. R. (1966). Projections of the lentiform nucleus in the monkey. *Brain Res.* 1, 3–42. doi: 10.1016/0006-8993(66)90103-X

Niv, Y., Edlund, J. A., Dayan, P., and O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* 32, 551–562. doi: 10.1523/JNEUROSCI.5498-10.2012

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454. doi: 10.1126/science.1094285

Owen, A. M. (2004). Cognitive dysfunction in Parkinson's disease: the role of frontostriatal circuitry. *Neuroscientist* 10, 525–537. doi: 10.1177/1073858404266776

Packard, M. G., and Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* 25, 563–593. doi: 10.1146/annurev.neuro.25.112701.142937

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., et al. (2012). Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* 76, 998–1009. doi: 10.1016/j.neuron.2012.10.017

Peelen, M. V., and Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *J. Neurophysiol.* 93, 603–608. doi: 10.1152/jn.00513.2004

Ploghaus, A., Tracey, I., Clare, S., Gati, J. S., Rawlins, J. N. P., and Matthews, P. M. (2000). Learning about pain: the neural substrate of the prediction error for aversive events. *Proc. Natl. Acad. Sci. U.S.A.* 97, 9281–9286.

Provost, J.-S., Hanganu, A., and Monchi, O. (2015). Neuroimaging studies of the striatum in cognition Part I: healthy individuals. *Front. Syst. Neurosci.* 9:140. doi: 10.3389/fnsys.2015.00140

Ramnani, N., Elliott, R., Athwal, B. S., and Passingham, R. E. (2004). Prediction error for free monetary reward in the human prefrontal cortex. *Neuroimage* 23, 777–786. doi: 10.1016/j.neuroimage.2004.07.028

Rief, W., and Glombiewski, J. A. (2016). Expectation-focused psychological interventions (EFPI). *Verhaltenstherapie* 26, 47–54.

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., and Thorwart, A. (2015). Expectancies as core feature of mental disorders. *Curr. Opin. Psychiatry* 28, 378–385. doi: 10.1097/YCO.0000000000000184

Rodriguez, P. F., Aron, A. R., and Poldrack, R. A. (2006). Ventral–striatal/nucleus–accumbens sensitivity to prediction errors during classification learning. *Hum. Brain Mapp.* 27, 306–313. doi: 10.1002/hbm.20186

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E. (2009). Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage* 45, 810–823. doi: 10.1016/j.neuroimage.2008.12.039

Sarinopoulos, I., Grupe, D. W., Mackiewicz, K. L., Herrington, J. D., Lor, M., Steege, E. E., et al. (2010). Uncertainty during anticipation modulates neural responses to aversion in human insula and amygdala. *Cereb. Cortex* 20, 929–940. doi: 10.1093/cercor/bhp155

Schiller, D., Levy, I., Niv, Y., LeDoux, J. E., and Phelps, E. A. (2008). From fear to safety and back: reversal of fear in the human brain. *J. Neurosci.* 28, 11517–11525. doi: 10.1523/JNEUROSCI.2265-08.2008

Schlerf, J., Ivry, R. B., and Diedrichsen, J. (2012). Encoding of sensory prediction errors in the human cerebellum. *J. Neurosci.* 32, 4913–4922. doi: 10.1523/JNEUROSCI.4504-11.2012

Schönberg, T., Daw, N. D., Joel, D., and O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J. Neurosci.* 27, 12860–12867. doi: 10.1523/JNEUROSCI.2496-07.2007

Schonberg, T., O'Doherty, J. P., Joel, D., Inzelberg, R., Segev, Y., and Daw, N. D. (2010). Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. *Neuroimage* 49, 772–781. doi: 10.1016/j.neuroimage.2009.08.011

Seger, C. A. (2006). The basal ganglia in human learning. *Neuroscientist* 12, 285–290. doi: 10.1177/1073858405285632

Sescousse, G., Caldú, X., Segura, B., and Dreher, J.-C. (2013). Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. *Neurosci. Biobehav. Rev.* 37, 681–696. doi: 10.1016/j.neubiorev.2013.02.002

Seymour, B., Daw, N., Dayan, P., Singer, T., and Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *J. Neurosci.* 27, 4826–4831. doi: 10.1523/JNEUROSCI.0400-07.2007

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., et al. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat. Neurosci.* 8, 1234–1240. doi: 10.1038/nn1527

Simmons, A., Matthews, S. C., Paulus, M. P., and Stein, M. B. (2008). Intolerance of uncertainty correlates with insula activation during affective ambiguity. *Neurosci. Lett.* 430, 92–97. doi: 10.1016/j.neulet.2007.10.030

Simmons, W. K., Avery, J. A., Barcalow, J. C., Bodurka, J., Drevets, W. C., and Bellgowan, P. (2013). Keeping the body in mind: Insula functional organization and functional connectivity integrate interoceptive, exteroceptive, and emotional awareness. *Hum. Brain Mapp.* 34, 2944–2958. doi: 10.1002/hbm.22113

Spoormaker, V. I., Andrade, K. C., Schröter, M. S., Sturm, A., Goya-Maldonado, R., Sämann, P. G., et al. (2011). The neural correlates of negative prediction error signaling in human fear conditioning. *Neuroimage* 54, 2250–2256. doi: 10.1016/j.neuroimage.2010.09.042

Takemura, H., Samejima, K., Vogels, R., Sakagami, M., and Okuda, J. (2011). Stimulus-dependent adjustment of reward prediction error in the midbrain. *PLoS ONE* 6:e28337. doi: 10.1371/journal.pone.0028337

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *J. Neurophysiol.* 95, 301–310. doi: 10.1152/jn.00762.2005

Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., and Fox, P. (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum. Brain Mapp.* 33, 1–13. doi: 10.1002/hbm.21186

Valentin, V. V., and O'Doherty, J. P. (2009). Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. *J. Neurophysiol.* 102, 3384–3391. doi: 10.1152/jn.91195.2008

Watanabe, N., Sakagami, M., and Haruno, M. (2013). Reward prediction error signal enhanced by striatum–amygdala interaction explains the acceleration of probabilistic reward learning by emotion. *J. Neurosci.* 33, 4487–4493. doi: 10.1523/JNEUROSCI.3400-12.2013

Wunderlich, K., Smittenaar, P., and Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron* 75, 418–424. doi: 10.1016/j.neuron.2012.03.042

Yanike, M., and Ferrera, V. P. (2014). Interpretive monitoring in the caudate nucleus. *eLife* 3:e03727. doi: 10.7554/eLife.03727

Zaki, J., Davis, J. I., and Ochsner, K. N. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *Neuroimage* 62, 493–499. doi: 10.1016/j.neuroimage.2012.05.012

Conflict of Interest Statement: 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.

Copyright © 2017 D'Astolfo and Rief. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

REVIEW
published: 22 February 2017
doi: 10.3389/fpsyg.2017.00244

The Dopamine Prediction Error: Contributions to Associative Models of Reward Learning

Helen M. Nasser[1], Donna J. Calu[1], Geoffrey Schoenbaum[1,2,3] and Melissa J. Sharpe[2,4]**

[1] Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, MD, USA, [2] Cellular Neurobiology Research Branch, National Institute on Drug Abuse Intramural Research Program, Baltimore, MD, USA, [3] Solomon H. Snyder Department of Neuroscience, Johns Hopkins University, Baltimore, MD, USA, [4] Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

Phasic activity of midbrain dopamine neurons is currently thought to encapsulate the prediction-error signal described in Sutton and Barto's (1981) model-free reinforcement learning algorithm. This phasic signal is thought to contain information about the quantitative value of reward, which transfers to the reward-predictive cue after learning. This is argued to endow the reward-predictive cue with the value inherent in the reward, motivating behavior toward cues signaling the presence of reward. Yet theoretical and empirical research has implicated prediction-error signaling in learning that extends far beyond a transfer of quantitative value to a reward-predictive cue. Here, we review the research which demonstrates the complexity of how dopaminergic prediction errors facilitate learning. After briefly discussing the literature demonstrating that phasic dopaminergic signals can act in the manner described by Sutton and Barto (1981), we consider how these signals may also influence attentional processing across multiple attentional systems in distinct brain circuits. Then, we discuss how prediction errors encode and promote the development of context-specific associations between cues and rewards. Finally, we consider recent evidence that shows dopaminergic activity contains information about causal relationships between cues and rewards that reflect information garnered from rich associative models of the world that can be adapted in the absence of direct experience. In discussing this research we hope to support the expansion of how dopaminergic prediction errors are thought to contribute to the learning process beyond the traditional concept of transferring quantitative value.

Keywords: prediction error, attention, associative learning, dopamine, model-based learning

OPEN ACCESS

Edited by:
Anna Thorwart,
University of Marburg, Germany

Reviewed by:
Jacqueline Gottlieb,
Yale School of Medicine, USA
Matt Lattal,
Oregon Health and Science University, USA
Ronald Keiflin,
Johns Hopkins University, USA

****Correspondence:***
Helen M. Nasser
helennasser2200@gmail.com
Melissa J. Sharpe
melissa.sharpe@nih.gov

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 October 2016*
Accepted: *07 February 2017*
Published: *22 February 2017*

Citation:
Nasser HM, Calu DJ, Schoenbaum G and Sharpe MJ (2017) The Dopamine Prediction Error: Contributions to Associative Models of Reward Learning. Front. Psychol. 8:244. doi: 10.3389/fpsyg.2017.00244

INTRODUCTION

The discovery that midbrain dopaminergic neurons exhibit a strong phasic response to an unexpected reward which subsequently transfers back to a cue which predicts its occurrence has been revolutionary for behavioral neuroscience (Schultz, 1997; Schultz et al., 1997). This was in part because this pattern of firing mimics the teaching signal predicted to underlie learning in models of reinforcement learning (Bush and Mosteller, 1951; Rescorla and Wagner, 1972; Mackintosh, 1975; Pearce and Hall, 1980; Sutton and Barto, 1981). The key concept in these learning models is that learning about reward-predictive cues is regulated by prediction error. When a subject

experiences a reward that they did not anticipate in the presence of a cue, a prediction error is elicited to drive learning so that the antecedent cue comes to motivate behavior directed toward the outcome. This prediction error is generally conceptualized as a quantitative discrepancy between the outcome expected when the cue was presented, and the outcome that was actually experienced (Bush and Mosteller, 1951; Rescorla and Wagner, 1972; Sutton and Barto, 1981). In essence, when an individual first encounters a cue followed by an unexpected reward, there is a large discrepancy between what is expected and what actually occurs, producing a large prediction error. However, when an individual learns that a particular cue reliably predicts a motivationally-significant event, there is little error as the discrepancy between what is expected and what actually occurred is diminished. Thus the prediction error functions to drive learning about reward-predictive cues and facilitate more accurate predictions about future rewards.

As the field now stands, phasic activity of midbrain dopamine neurons is considered to represent the prediction error that drives learning as described by the Sutton and Barto (1981) model-free reinforcement learning algorithm. This algorithm explicitly conceptualizes the discrepancies between the expected and delivered outcome as reflecting differences in predicted value, and computes the resultant prediction errors over consecutive time steps during a trial. As a result, the value signal usually produced by reward transfers temporally back to events that reliably precede reward delivery. This effectively endows a cue that predicts reward with the value inherent in the reward itself, rather than just registering when the reward has occurred. In this manner, Sutton and Barto's (1981) model-free reinforcement learning algorithm explicitly states that the quantitative value inherent in reward transfers back to the antecedent cue predicting its delivery. That is, the predictive cue becomes endowed with the scalar value of the reward rather than explicitly predicting the identity of the outcome which follows cue presentation.

However, thinking about firing from dopaminergic neurons as reflecting a quantitative value signal is limited and does not allow this phasic signal to influence many other complex forms of learning. Firstly, we do not associate all cues with the rewards that they precede. Rather, we select particular cues to learn about on the basis of how well they have predicted that particular reward, or any reward in the past. Such a tendency is encapsulated in models of selective attention in associative theory (Mackintosh, 1975; Pearce and Hall, 1980), where attention directed toward a cue will vary by virtue of its ability to predict reward in the past. But in these models of selective attention, attentional signals are critically influenced by prediction error. That is, the prediction-error signal explicitly informs the change in attention directed toward a cue. Secondly, humans and animals are also capable of inferring associations between cues and rewards in the absence of direct experience. For example, if a cue has been established as predictive of a particular reward and that reward is then devalued outside of the experimental context, the subject will change how they respond to the cue on their next encounter with the cue. This is despite never directly experiencing the now devalued outcome in the presence of the cue. Such learning is typically referred to as 'model-based' and is not under the control of the Sutton and Barto (1981) error signal which relies on cached values drawn from direct experiences with cues and outcomes (Dickinson and Balleine, 2002; Berridge, 2012; Dayan and Berridge, 2014). However, recent evidence has begun to suggest that phasic dopamine signals in the midbrain may incorporate model-based information (Bromberg-Martin et al., 2010c; Daw et al., 2011; Hong and Hikosaka, 2011; Aitken et al., 2016; Cone et al., 2016; Sadacca et al., 2016). Such evidence suggests that the dopaminergic error signal may not exist completely apart from these other more complex learning mechanisms.

Here we review empirical studies that challenge and expand on how the dopamine prediction error incorporates and influences learning at associative and circuit levels. In doing so, we will first briefly review the neural correlates of the bidirectional prediction-error signal contained in phasic activity in midbrain dopamine neurons. Then, we will move onto a discussion of how this signal may support a change in attention across multiple attentional systems in distinct brain circuits. Finally, we will review recent evidence that suggests the information contained in the phasic dopamine signal extends beyond that conceptualized by a model-free account. In particular, midbrain dopamine signals appear to reflect information about causal relationships between cues and outcomes in a manner that extends beyond simply encoding the value of a reward predicted by a cue. Such research expands the currently narrow view of how phasic dopamine activity can influence the learning process.

REWARD PREDICTION ERROR SIGNALS

At the core of the Sutton and Barto (1981) model-free reinforcement learning algorithm is the concept that prediction error drives learning about cues and the outcomes they predict. That is, if an individual experiences an outcome they did not expect when a cue is presented, a teaching signal will be elicited to update expectations and reduce that prediction error. As a reward in this context is conceptualized as containing an inherent quantitative value, it is this quantitative value that is thought to be transferred to the predictive cue. Effectively, this is argued to endow that predictive cue with the scalar expectation of the upcoming reward. Furthermore, this algorithm proposes that prediction error is bidirectional. Thus, it can drive increases or decreases in learning via signaling a positive or negative prediction error, respectively. A positive prediction error will be elicited when a cue predicts a reward that was more valuable than expected. Here, this signal will act to increase the value attributed to the antecedent cue. However, if an outcome is less valuable than expected on the basis of the expectation elicited by the antecedent cue, a negative prediction error will be elicited and the prediction-error signal will act to reduce the value held by the cue. Essentially, this allows the prediction-error teaching signal to regulate both increases and decreases in the value attributed to predictive cues as a function of the quantitative difference in the reward expected relative to that delivered.

Electrophysiological studies in rodents and non-human primates have demonstrated very convincingly that phasic dopaminergic activity can correlate with the prediction error

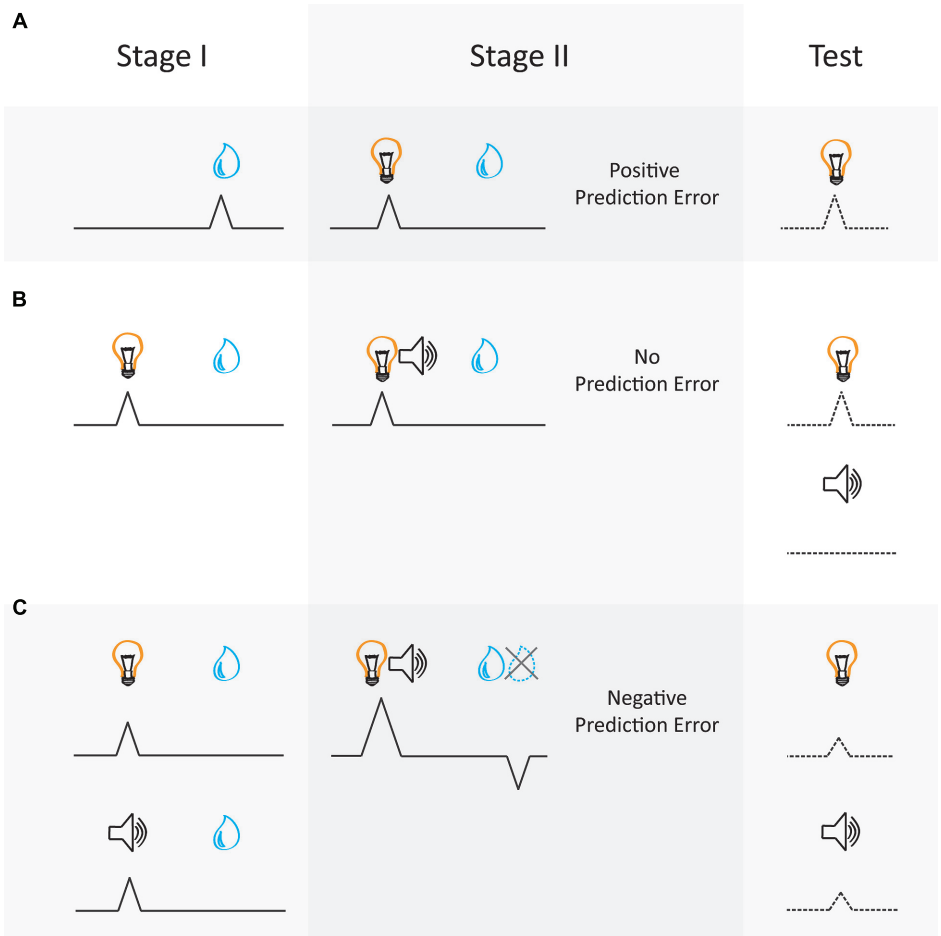


FIGURE 1 | Dopamine neural correlates follow the laws of prediction error as formalized in Sutton and Barto's (1981) model-free reinforcement learning algorithm. According to Sutton and Barto (1981) prediction errors are generated by the quantitative difference between the value of the reward delivered and the value attributed to the reward-predictive cue. Phasic activity of dopaminergic neurons in the VTA can be interpreted as conforming with these predictions. This schematic diagram represents dopamine activity during different Pavlovian conditioning paradigms. Activity is represented by the black lines aligned to a cue (e.g., light or tone, on the left) and reward (e.g., a juice drop, on the right). The dotted lines refer to changes in associative strength. **(A)** Pavlovian conditioning: A reward elicits a positive prediction error when the reward is unexpected, thus an error is made in prediction (Stage I). Dopamine neurons exhibit firing upon the reward delivery. However, with repeated cue-reward pairings this dopamine signal transfers to the reward-predictive cue and diminishes to the reward (Stage II) (Hollerman and Schultz, 1998). As this cue is now predictive of reward, there is a reduction in prediction error at the time of reward and motivated behavior directed toward the predictive cue increases. **(B)** Blocking: a critical aspect of the Sutton and Barto (1981) model is that the learning (or value) about the reward must be shared amongst all present cues. This is referred to in learning theory as a *summed-error term* (Rescorla and Wagner, 1972). This concept is well illustrated by the blocking phenomenon. For example, during Stage I a light cue is trained to predict reward and with training comes to elicit a dopamine signal (Waelti et al., 2001). When a second auditory cue (tone) is presented simultaneously with the light cue and the same quantity of reward is delivered during Stage II, no prediction error is elicited as the reward is already expected and no dopamine signal is exhibited. Behaviorally, learning about the novel tone cue is said to be blocked, and when the cues are presented alone at Test the light cue maintains associative strength but the blocked tone cue does not gain any associative strength. **(C)** Over-expectation: Two different cues (light and tone) that have been separately trained to predict a particular quantity of reward come to each elicit a dopamine prediction-error signal after multiple cue-reward pairings in Stage I. During Stage II, the two cues are then presented as a simultaneous compound, followed by reward given to each trial type during Stage I. This generates a negative prediction error, as the reward is less than the summed expectation of each cue. In this example dopamine signaling is suppressed in response to the over-expected reward not being delivered. This negative prediction error drives a reduction in associative strength so that both cues lose half their associative value when presented alone at Test, assuming these cues are matched for salience (e.g., Chang et al., 2016).

contained in Sutton and Barto's (1981) model (**Figure 1**). These neurons show a phasic increase in activity when an unexpected reward is delivered (Ljungberg et al., 1992; Mirenowicz and Schultz, 1994, 1996) or a reward is delivered that was better than expected (Bayer and Glimcher, 2005) (**Figure 1A**). Further, the magnitude of phasic activity correlates with the size of the unexpected reward (Hollerman and Schultz, 1998; Fiorillo et al., 2003; Roesch et al., 2007; Stauffer et al., 2014; Eshel et al., 2016) in a manner that reflects the value of the reward (Lak et al., 2014), value of the future action (Morris et al., 2006) or value of the choice (Roesch et al., 2007), as assessed by the agent's approach behavior toward to reward-predictive cue. That is, the firing of dopamine neurons changes in response to unexpected rewards or reward-predictive cues in a manner that appears to reflect

the subjective value of those rewards. Additionally, the firing of dopaminergic neurons in the midbrain is suppressed when an expected reward is omitted or is worse than expected (Tobler et al., 2003; Brischoux et al., 2009; Matsumoto and Hikosaka, 2009; Lammel et al., 2011, 2014; Cohen et al., 2012). Finally, dopamine neurons also show a slow reduction of firing to the reward over successive cue-reward pairings as the cue comes to reliably predict the reward (Hollerman and Schultz, 1998). That is, the now expected reward elicits minimal phasic excitation when it is presented after the cue, where this activity instead shifts to presentation of the cue itself (see **Figure 1A**). Thus there is a wealth of empirical evidence that can be interpreted as supporting the idea that dopaminergic prediction-error signals comply with those predicted by Sutton and Barto's (1981) model-free reinforcement learning algorithm.

Another critical aspect of Sutton and Barto's (1981) model is that associative strength (or value) afforded by the reward must be shared amongst all present cues, referred to as a *summed-error term*. The presence of this summed-error term allowed earlier models (Rescorla and Wagner, 1972) to account for circumstances when cues are presented simultaneously and compete to become associated with the same outcome, as demonstrated in the blocking procedure (see **Figure 1B**). In one example (Waelti et al., 2001), monkeys first received presentations of cue A paired with a juice reward. In the second phase of training, novel cue X was introduced and presented simultaneously with cue A to form a compound AX, where presentation of cue AX was followed with the same juice reward as the first stage of training. During this second phase, monkeys also received a completely novel compound BY followed by the same juice reward. Here, as cue A had already become predictive of reward, there was no error in prediction when compound AX was presented and no associative strength accrued to cue X. On the other hand, as cue BY had never been paired with reward, both cues gained associative strength, sharing the value inherent in the juice reward. Thus when monkeys were tested with cue X and Y they responded more to cue Y as reward was only expected when cue Y was presented. This blocking effect illustrates how prediction error regulates learning by prioritizing cues that have already come to predict reward (Kamin, 1968), allocating less value to a novel cue which does not provide additional information about reward delivery. Thus, prediction errors regulate learning in a manner that produces causal relationships between a cue and the outcome it predicts.

Importantly, midbrain dopaminergic neurons also adhere to the principal of a summed-error term inherent in these models. Specifically, in the blocking design illustrated above (see **Figure 1B**), Waelti et al. (2001) recorded putative dopaminergic neurons during this task. As previously demonstrated dopaminergic neurons increased firing to cue A during the initial phase of training. Then, across the second phase of training dopaminergic neurons maintained similar firing rates to presentations of compound cue AX. Further, dopamine neurons also increased firing rate to the novel compound cue BY across this phase. Critically, in a non-reinforced test where cue X and Y were presented individually without reward, dopaminergic neurons showed robust phasic responses toward cue Y but no response to the blocked cue X, mimicking the behavioral response seen in the blocking paradigm. As cue X and Y were matched for physical salience and paired with an equivalent reward, any difference in the dopaminergic response to these cues could only be attributed to a difference in the summed prediction error, in line with that described by Sutton and Barto (1981).

Until very recently evidence suggesting that phasic activity in midbrain dopamine neurons mimics the scalar prediction error described in Sutton and Barto (1981) has been largely, if not entirely, correlative (Schultz et al., 1997; Roesch et al., 2007; Niv and Schoenbaum, 2008; Iordanova, 2009; Keiflin and Janak, 2015; Holland and Schiffino, 2016; Schultz, 2016). This is because it was difficult to directly manipulate dopamine neurons with the temporal precision and specificity required to directly test this hypothesis. However, the combination of a temporally specific optogenetic approach in addition to the cell type specific transgenic rodent lines has made it easier to manipulate dopamine neurons in a causal manner (Margolis et al., 2006; Lammel et al., 2008; Tsai et al., 2009; Witten et al., 2011; Cohen et al., 2012). This has been hugely advantageous to the study of how prediction-error signals causally influence the learning process. Using transgenic animals expressing Cre recombinase under the control of tyrosine hydroxylase promoter (i.e., Th::Cre lines), a precursor enzyme for dopamine, Cre-dependent viral-vectors injected in to the midbrain can be used to induce expression of the light-sensitive channelrhodopsin-2 (ChR2) or halorhodopsin (NpHR) to selectively activate or inhibit neurons expressing tyrosine hydroxylase (TH+ neurons), respectively. This has afforded neuroscientists the capacity to manipulate dopaminergic neurons in a temporally specific manner that mimics positive or negative prediction errors and assess their causal contribution to the learning process (Steinberg et al., 2013; Chang et al., 2016; Stauffer et al., 2016).

Using this technique, Steinberg et al. (2013) have causally demonstrated that stimulation of dopaminergic neurons in the midbrain can mimic a positive prediction error to drive learning. Steinberg et al. (2013) injected TH-Cre rats with ChR2 in the ventral tegmental area (VTA) and implanted optical fibers aimed at VTA. This allowed phasic stimulation of TH+ neurons in the VTA to mimic the phasic activity typically seen with an unexpected reward and drive excitatory learning. In order to test that these signals do in fact drive learning about reward-predictive cues, they used a blocking procedure, similar to that described above (Waelti et al., 2001; **Figure 1B**). Rats were first presented with cue A that signaled food reward. In a second phase of training, compound cue AX was paired with the same reward. No prediction-error signal should be elicited by the compound cue AX when the reward was presented in the second phase. Therefore, rats would exhibit little learning about cue X as the reward had already been predicted by cue A during training in the first phase of learning. When Steinberg et al. (2013) activated TH+ neurons to artificially mimic a positive prediction error during reward receipt following presentation of the compound cue AX, they found an increase in responding to the usually blocked cue, X, in the subsequent probe test. This

result suggests that activating dopaminergic neurons in the VTA mimics a positive prediction error to causally drive learning about the usually blocked cue, X.

If dopamine neurons truly reflect bidirectional prediction errors, it would be expected that briefly silencing their activity would produce a negative prediction error and drive down the ability of a cue to predict reward. In order to determine whether silencing dopaminergic neurons in the VTA could function as negative prediction errors in this manner, Chang et al. (2016) briefly silenced these neurons during a modified version of an over-expectation task (for a simplified illustration see **Figure 1C**). In the standard over-expectation task, the first phase of learning in over-expectation requires that rats learn about two cues (A and B) that independently predict the same magnitude of reward (e.g., one food pellet). During a second phase of learning these two cues are presented as compound AB followed by the same reward. Because cues A and B independently predict the same magnitude of reward, when AB is presented in compound, rats expect delivery of twice the amount of reward (e.g., two food pellets). As rats only receive one food pellet, a negative prediction error is elicited and the associative strength of both cues A and B decreases. However, in a modified version of the over-expectation task, Chang et al. (2016) presented rats with the compound cue AB in the second stage of learning with the expected two food pellets. This change effectively blocks over-expectation from occurring. Against this backdrop, they briefly suppressed TH+ neurons in the VTA during presentation of the reward in AB compound phase of learning. This manipulation decreased the ability of cues A and B to elicit a motivational response in the following probe test, just like what would usually be seen in the traditional over-expectation procedure. Thus Chang et al. (2016) found that transiently suppressing firing of TH+ neurons was sufficient to mimic a negative prediction error. Together, these studies confirm that phasic dopamine can serve as a bidirectional prediction error to causally drive learning.

It is worth briefly noting here that the blocking effect described above has been interpreted as reflecting a performance deficit rather than the result of less learning accruing to the blocked cue X (Miller and Matzel, 1988; Arcediano et al., 2004). According to the comparator hypothesis (Miller and Matzel, 1988), responding to a conditioned cue is in part the result of the strength of the direct association between this cue and the outcome. However, it is also inversely related to the associative strength of any other cue that is presented within a session (i.e., the comparator cue). In this sense, reduced responding to the blocked cue X at test is argued to be the result of increased associative strength that has already accrued to the comparator cue A during the initial phase of conditioning. The evidence in favor of a performance account of blocking is contradictory (Miller and Matzel, 1988; Blaisdell et al., 1999), however, in some instances it has been shown that responding to the blocked cue, X, can be recovered by massive extinction of the comparator cue A which is consistent with the comparative hypothesis (Blaisdell et al., 1999). This research may have consequences for how we interpret VTA DA signals during the blocking task. Specifically, it raises the possibility that the reduced response of dopamine neurons to the blocked cue during the extinction test may reflect the signal used for responding to the blocked cue as predicted by the performance account, rather than the direct association between the blocked cue and the outcome. In this manner, this signal could comprise the quantitative combination of the direct association between the blocked cue X and the outcome, as well as the inverse of the associative strength of the comparator cue A. According to this interpretation, it would not constitute a teaching signal driving learning but rather a signal which reflects this comparative process to produce the reduced response. However, the causal data showing that phasic stimulation of VTA dopamine neurons unblocks learning about the blocked cue X, which results in an increased response to the cue in a subsequent extinction test without stimulation (Steinberg et al., 2013), suggests that these error signals act to causally influence the learning process rather than simply reflecting a comparator signal used for performance.

ATTENTION

The VTA resides within a rich neural circuit, sending and receiving dense projections from subcortical and cortical regions. Thus it is not surprising that prediction-error signaling in VTA has important and wide-reaching consequences for reward processing across distributed brain reward circuits. For example, prediction-error signaling in VTA influence downstream processing of attention paid toward cues (Corlett et al., 2007; Berridge, 2012; Roesch et al., 2012; Holland and Schiffino, 2016). Interestingly, the manner in which VTA signaling appears to do this has again been predicted by associative models many years before neuroscientists were able to examine these circuits in the way we can today. More interesting still, the mechanisms by which VTA signaling may facilitate attentional processing are diverse and mirrors the controversy in the reinforcement learning literature.

Specifically, a contradiction which has confused undergraduate psychology students for decades is the opposing predictions made by the two dominant attentional theories in associative learning, namely the Mackintosh (1975) and Pearce and Hall (1980) models. On the one hand, Mackintosh's (1975) model of attention argues that attention will be paid to cues in the environment that are the *best predictors* of a motivationally significant event. Yet, the Pearce and Hall (1980) model of attention predicts the exact opposite- we should attend to cues when we are *uncertain of their consequence*. Indeed, there is strong evidence in humans and other animals for both of these attentional models which suggests that these contradictory attentional processes both exist and in fact contribute to attentional processing.

But each of the attentional strategies proposed by Mackintosh (1975) and Pearce and Hall (1980) models may be beneficial in different circumstances. Consider a situation where we have many cues which predict reward with differing accuracy. Here, it is more efficient to devote attention toward cues that are the best predictors to maximize reward, in line with a Mackintosh (1975) process. However, in a scenario where one or a few cues predict reward it is not always beneficial to devote a lot of attention to a cue that always predicts reward when it is not in direct

competition with another cue. Effectively, you do not need to pay a lot of attention to a cue when it is the only one available and attention does not need to bias action selection, in line with the Pearce and Hall (1980) model of attention. Rather, it becomes more important to detect changes in the contingency between a cue and reward to update our knowledge of these relationships.

Evidence for a view where different scenarios recruit different attentional processes is supported by the fact that findings consistent with either Mackintosh (1975) or Pearce and Hall (1980) models tend to be found using different experimental parameters. Individuals are generally found to attend to the best predictors of reward when parameters promote high cue competition (Mackintosh, 1965, 1973, 1976; McLaren and Mackintosh, 2002; Le Pelley et al., 2011, 2013) whereas effects suggesting individuals attend more to inconsistent predictors are generally found in cases where one or few cues are available (Hall and Pearce, 1979; Wilson et al., 1992; Griffiths et al., 2011; Esber et al., 2012). In fact, recent models of associative learning have formalized this concept to predict how attention will change across learning under these different circumstances, via hybrid models (LePelley and McLaren, 2004; Pearce and Mackintosh, 2010) or models that reconcile the roles of predictiveness and uncertainty (Esber and Haselgrove, 2011).

Important to the current discussion is that models of reinforcement learning utilize prediction errors in two ways (Pearce and Hall, 1980; Mackintosh, 1975). Firstly, prediction-error signaling regulates the amount of learning that can occur on any single cue-reward pairing. That is, the magnitude of the difference between the expected and experienced reward will determine how much learning can accrue to the cue in subsequent trials. However, prediction errors are also argued to regulate the change in attention devoted to that cue, which will dictate the rate of learning and, therefore, which cues are learnt about. In Mackintosh's (1975) model, attention declines to cues that result in larger prediction errors and are, therefore, poor predictors of reward. Here, attention increases toward cues which results in a smaller prediction error relative to other present cues. In direct contrast, the Pearce and Hall (1980) model posits that attention is maintained to a cue that produces larger prediction errors. According to Pearce and Hall (1980), attention decreases when prediction errors are small, consequently well-established predictors will receive less attention.

The neural evidence also favors the presence of both these dissociable attentional processes. Specifically, evidence suggests that a Mackintosh-like (Mackintosh, 1975) attentional process occurs in the prelimbic cortex (PL) in the medial prefrontal cortex (mPFC) (Sharpe and Killcross, 2014, 2015), while neural activity in basolateral complex of the amygdala (BLA) reflects a Pearce and Hall (1980) signal (Roesch et al., 2010, 2012; Esber et al., 2012; Esber and Holland, 2014). Of course, such opposing attentional processes do not exist in isolation. It is well-established that VTA sends out dense projections to both the PL and BLA, providing a plausible circuit through which prediction-error signaling could influence attentional signals in these regions (see **Figure 2**). The presence of these dissociable neural circuits strengthens recent attempts to build models of associative learning which allows prediction error to influence attentional processing in these different ways (LePelley and McLaren, 2004; Pearce and Mackintosh, 2010; Esber and Haselgrove, 2011). That is, the neural evidence supports the idea that prediction error can regulate not only the amount of learning available on any one trial but also to influence different types of attentional processing in distinct circuits. In this section, we will examine the neural evidence for each of these systems alone and will then review recent attempts at a reconciliation between these attentional processes.

As a brief note here, we would acknowledge that we have focused on reviewing the literature which conceptualizes attention as a modulator of learning rates. That is, we have focused on models in which attention directly acts to regulate the amount of learning that is attributed toward a particular cue on any one trial. Conceptualizing attention in this manner has become common place within the associative learning literature, predominantly driven by studies utilizing rodents (but see: Le Pelley et al., 2011, 2016). However, there is a wealth of literature on attention which conceptualizes attention in other ways, mainly driven by studies in humans and non-human primates. For example, attention may also be conceptualized as modulating the bottom-up sensory processing of cues, or as influencing activation of cue-response associations (to name just a few; Miller and Cohen, 2001; Hickey et al., 2006). These mechanisms focus on how cues are processed relative to other present cues or how cues can influence the ability to elicit an associated response, but not the ultimate amount of learning that accrues to the cue itself. While the relationship between attention and behavior is likely the same across both sets of definitions- where increases in attention act to increase behavior directed toward a cue, and decreases in attention the reverse- there are significant differences in *how* attention is hypothesized to influence learning and/or behavior. Given this, it is likely that future integration of these fields would likely be fruitful in understanding attentional processing across species (see e.g., Hickey et al., 2006, 2011, 2015; Jovancevic et al., 2006; Hare et al., 2011; Hickey and Theeuwes, 2011; Lim et al., 2011; Gottlieb, 2012; Gottlieb et al., 2014; Theeuwes, 2013; Tommasi et al., 2015; Wilschut et al., 2015 for a more comprehensive review on these attentional theories).

Pearce and Hall (1980) Model of Attention

A sub-nucleus of the amygdala complex, the BLA, is a region that receives extensive dopaminergic input from midbrain dopamine neurons (Swanson, 1982) and shows increases in neural activity when an unexpected event occurs whether it is rewarding or aversive (Belova et al., 2007, 2008; Herry et al., 2007; Roesch et al., 2010; Tye et al., 2010; Li et al., 2011; Beyeler et al., 2016). Notably these signals seem to conform closely to what is predicted for a Pearce and Hall (1980) attentional signal. Specifically, Roesch et al. (2010) recorded neurons in the BLA during a task in which expectations were repeatedly violated. Here, rats were trained to enter a food well after two odors were presented. One of these odors predicted that the right well would be reinforced and the other predicted that the left well would be reinforced. At the beginning of each training block, the timing and size of rewards delivered in these wells were manipulated to either increase or decrease the value of the reward delivered at each

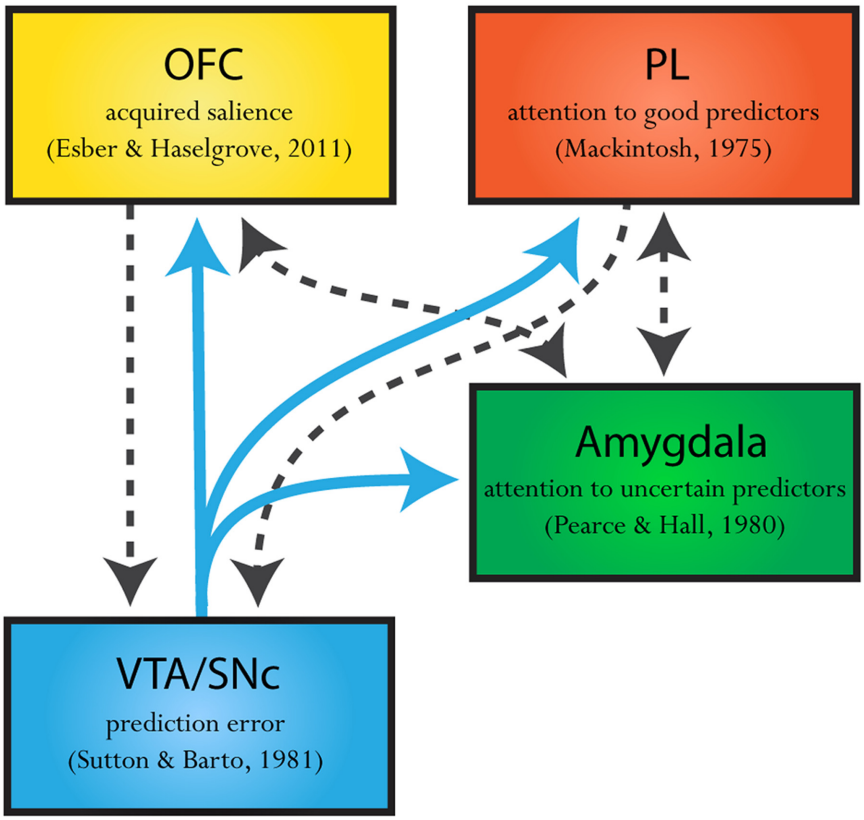


FIGURE 2 | Theorized neural circuitry of a dopaminergic prediction error projections informing attentional processes. Dopamine prediction-error signaling (blue solid arrows) regulates the amount of learning that can occur on any one cue-reward pairing. The midbrain, ventral tegmental area and substantia nigra (VTA/SNc) sends out dense dopaminergic projections to the amygdala, prelimbic cortex (PL) and orbitofrontal cortex (OFC). The amygdala, PL, and OFC have all been implicated in distinct attentional processes. Specifically, the amygdala has been implicated in attending toward cues which inconsistently predict an outcome, uncertain predictors, in line with a Pearce and Hall (1980) mechanism. On the other hand, the PL region of the mPFC has been implicated in promoting attention toward cues which are good predictors of an outcome relative to other reward-predictive cues, as predicted by Mackintosh's (1975) model of attention. Further still, neural correlates in the OFC appear to respond to cues which are good and uncertain predictors, where activity in a proportion of these neurons is high for cues that are consistent predictors of reward yet higher still for cues that are inconsistent predictors of reward (Ogawa et al., 2013) as predicted by Esber and Haselgrove (2011). Suggesting a role for the OFC in modulating acquired salience. The commonality of these three distinct attentional theories is that they all propose that prediction errors influence how much attention will be paid toward a cue. The presence of this neural circuit (illustrated above), where the VTA/SNc sends distinct projections to these three attentional regions provides a plausible circuit whereby prediction errors could influence attentional processing in different ways. Furthermore the PL, OFC, and amygdala are interconnected in such ways that these regions may integrate attentional information as described in two-process models (black dashed arrows; see McDonald, 1987, 1991; Vázquez-Borsetti et al., 2009).

well. Roesch et al. (2010) found that a population of neurons in the BLA responded similarly to both upshifts and downshifts of reward value. Specifically, these neurons increased their firing rate when expectations were violated, regardless of whether they constituted decreases or increases in reward value. This *unsigned* or unidirectional error signals are reminiscent of that described by the Pearce and Hall (1980) model of attention, whereby attention is enhanced by means of an absolute value prediction error. In line with an attentional interpretation, this neural signal was integrated across trials and correlated with greater levels of orienting toward the predictive cues after changes in reward, where orienting constitutes a reliable measure of overt attention in the associative learning literature. Functional inactivation of the BLA disrupted changes in orienting behavior and reduced learning to respond to changes in the reward. The findings from this study suggested that the BLA is critical in driving attention for learning according to a Pearce and Hall (1980) mechanism.

Notably, Esber et al. (2012) further demonstrated that the ability of BLA neurons to exhibit this Pearce-Hall signal is dependent on dopaminergic input from the VTA. Specifically, Esber et al. (2012) recorded neurons in the BLA of rats with ipsilateral sham or 6-hydroxydopamine (6-OHDA) lesions of the VTA during the choice task described above (Roesch et al., 2010). They found that neurons in the BLA of intact rats again showed this characteristic increase in activity to either upshift or downshifts in reward value in this task. However, BLA neurons in 6-OHDA-lesioned rats failed to show this attentional signal. Interestingly, despite the deficit in attentional signaling, neurons of lesioned rats still exhibited a sensitivity to value *per se*. That is, neurons in the BLA of lesioned rats continued to respond

more to cues predicting high magnitude of reward and less to those predicting lower amounts of reward. This demonstrated that dopaminergic activity in the VTA is necessary for neurons in the BLA to exhibit this unsigned prediction error but not for the ability of this region to encode other characteristic neuronal signals. Of course, while 6-OHDA lesions suppress phasic dopamine signaling, these lesions also suppress tonic dopamine signaling in the VTA. Thus, while it is clear that dopaminergic input appears to be necessary for neurons in the BLA to exhibit unsigned attentional signal in a manner described by the Pearce and Hall (1980) model, future research is necessary to confirm that relay of phasic VTA DA prediction-error signals produce an increase in attention toward a cue when expectations have been violated.

Interestingly, the central nucleus of the amygdala (CeA), another sub-nucleus of the amygdala complex, has also been implicated in attentional processes predicted by the Pearce and Hall (1980) model. Using a serial-conditioning task designed by Wilson et al. (1992), lesions of the CeA disrupted surprised-induced increments in attention (Holland and Gallagher, 1993). Here, two cues were presented as a serial compound, whereby a light consistently predicted presentation of a tone. On half of the trials, the serial compound was followed by reward. According to the Pearce and Hall (1980) model, as the light consistently predicted the tone, the attention to the light should be low. In a second phase, one group of rats continued this training. However, another group now only received the light prior to the tone on reinforced trials. On non-reinforced trials, the light was presented alone. That is, in this second group the light no longer consistently predicted the tone. According the Pearce and Hall (1980) this surprising omission of the tone should increase attention paid to light. Consistent with this prediction, sham-lesioned rats who received the surprising omission of the tone later showed faster acquisition of responding to the light when it was paired with a novel outcome than sham-lesioned rats that had consistent training. This showed that attention to the light increased as a consequence of the omission of the tone which facilitated later learning about the light. However, rats with lesions of the CeA failed to show this faster rate of learning as a consequence of the surprising omission of the tone. This demonstrated that the CeA is necessary for surprise-induced increments in attention, in line with predictions made by the Pearce and Hall (1980) model of attention.

The role for the CeA in surprise-induced increments in attention is not dissimilar from the attentional role attributed to the BLA. That is, both regions have been implicated in increases in attention as a result of the violation of expectancies in line with the Pearce and Hall (1980) model. However, while this attentional process in BLA appears to be the product of direct dopaminergic projections from the VTA, the CeA does not receive this input (Pitkanen et al., 2000). Rather, the CeA receives projections from the substantia nigra (SNc) that appears to facilitate this attentional process. Specifically, Holland and Gallagher (1993) demonstrated that disconnection of the SNc and CeA using ibotenic acid lesions of CeA in one hemisphere and 6-OHDA lesions of SNc in the opposite hemisphere prevented increasing attention to the light cue when it no longer consistently predicted the tone in the serial-conditioning task described above (Lee et al., 2006). This demonstrates that it is dopaminergic input from the SNc that facilitates attentional processing in the CeA, rather than from the VTA, as appears to be the case in the BLA. This anatomical difference invites the possibility that the attentional processes taking place in these regions are fundamentally different. This possibility is supported by the finding that lesions of the CeA also interfere with the basic acquisition of a conditioned orienting response to a reward-predictive cue, whereas BLA lesions do not (Holland and Gallagher, 1993, 1999). This has led to the argument that CeA drives behavioral changes resulting from changes in attention (Holland and Gallagher, 1999; Holland et al., 2001). Thus, dopamine projections from the SNc to CeA may function to produce overt behavioral changes in attention to influence rates of learning rather than modulating the rate at which a cue becomes associated with an outcome *per se,* which may be a point of difference from attentional processing which takes place in the BLA.

Mackintosh (1975) Model of Attention

In contrast to the role of the CeA and BLA in an attentional process implicated in the Pearce and Hall (1980) model, inhibition of activity in the rodent mPFC has been causally demonstrated to produce deficits in modulating attention toward cues in a manner akin to that described by Mackintosh's (1975) theory of attention (Sharpe and Killcross, 2014, 2015). As would be expected from a region modulating attention according to a Mackintosh (1975) attentional process, lesions or inactivation of the mPFC produce deficits in tasks that promote high competition between multiple cues. The classic finding is that mPFC lesions produce impairments in extradimensional set shifting, where subjects have to attend toward a set of cues that are established as predictive of reward and disregard other present, but irrelevant, cues (Birrell and Brown, 2000). Such effects have more recently been attributed to the PL region of the mPFC, where a role for this region in attention can now be explicitly dissociated from a role in error correction (Sharpe and Killcross, 2014, 2015). For example, PL lesions do not disrupt expression of the blocking effect but selectively impair the ability to stop attending toward the redundant blocked cue (Sharpe and Killcross, 2014). Here, rats received PL lesions prior to a typical blocking paradigm. In stage I of this task, rats received pairings of cue A with reward. In stage II cue A was paired with novel cue B and the same magnitude of reward. In this same stage, rats were also presented with a novel compound CD and the same reward. PL lesions did not affect blocking to cue B relative to cue D, demonstrating an intact error-correction process dependent on prediction-error signaling in the VTA. However, after the blocking procedure these same animals were presented with the blocked cue B and then presented with reward. In line with a Mackintosh attentional process, sham-lesioned rats demonstrated slow learning about cue B, suggesting attention had declined toward this cue as it was previously a poor predictor of the outcome. However, rats with PL lesions exhibited faster learning about this cue suggesting they had not down-regulated attention toward blocked cue B.

This demonstrates that the PL cortex is necessary to direct a preferential degree of attention toward predictive cues while not being necessary to allow learning to be regulated by prediction error *per se*.

Interestingly, VTA sends a particularly dense projection to the PL region (Bentivoglio and Morelli, 2005; Björklund and Dunnett, 2007). While the causal influence of these signals on attentional processing are lacking and constitute an interesting direction for future research, there is considerable evidence that phasic firing in VTA dopamine neurons directly affects neurons in the mPFC (Niki and Watanabe, 1979; Tzschentke and Schmidt, 2000; Rushworth et al., 2011). For example, electrophysiological studies have demonstrated that burst stimulation of VTA promotes prolonged depolarization of mPFC pyramidal neurons, constituting a change to an 'up state' where the membrane potential of neurons in this area is brought close to firing threshold (Lewis and O'Donnell, 2000). Such research may suggest that phasic firing in VTA may act to enable plasticity in mPFC circuits, where firing rates tune to cues which are good predictors of an outcome (Gruber et al., 2010). In line with this, evidence from electrophysiology (Niki and Watanabe, 1979) and functional magnetic resonance imaging (fMRI) studies (Rushworth et al., 2011), have shown that activity in mPFC encodes both the value of the upcoming rewards predicted by cue presentation as well as a depression in activity during the omission of an expected reward. This is distinct from activity seen in the BLA which, as discussed above, exhibits a general increase in firing in response to both delivery and omission of expected reward. While the mechanism by which burst firing in VTA dopamine neurons influence attentional processing in PL cortex remains to be clarified, the ability of phasic responses to influence activity in PL cortex suggest that prediction errors in VTA may influence activity in the PL cortex to produce an attentional signal in line with that predicted by Mackintosh's (1975) model of attention and dissociable from that seen in other regions of the brain.

It is worth noting here that the neural signal predicted by Mackintosh's (1975) model of selective attention is not as simple as an increase in responding to cues which are predictive of reinforcement. Indeed, many regions of the brain show increases in activity to predictive cues. The uniqueness of Mackintosh's (1975) predicted attentional signal is perhaps best illustrated by the model's predictions in times of cue competition. Take, for example, the overshadowing paradigm, whereby an audio–visual compound is presented with reward. If this compound cue differs in intrinsic salience, after the first few trials associative strength will decrease toward the less intrinsically salient element of the compound (a dim visual cue) as the more intrinsically salient element of the compound (a loud auditory cue) accrues associative strength more quickly, and this overshadows the less intrinsically salient element. Unlike most models of reinforcement learning (Rescorla and Wagner, 1972; Pearce and Hall, 1980; Sutton and Barto, 1981), Mackintosh's (1975) model does not use the summed-error term developed in the Rescorla and Wagner (1972) model, later adapted by Sutton and Barto (1981). Instead, learning to predict an outcome need not be shared by all present cues. Mackintosh's (1975) model uses attentional change to explain the decrement in learning when multiple cues of different intrinsic salience predict the same outcome. More formally, the change in a cue's associative strength is based on that individual cue's prediction error. Thus the less intrinsically salient cue is learnt about more slowly and is, therefore, a less reliable predictor of reward and learning about this cue stops. In line with a role for the PL cortex in a Mackintosh (1975) attentional process, inactivation of the PL cortex specifically impairs overshadowing of the less intrinsically salient visual cue paired with a shock in a procedure that promotes this form of overshadowing (Sharpe and Killcross, 2015).

The presence of an individual-error term in the Mackintosh (1975) model has consequences for the nature of the attentional signal that may expected in neural regions contributing to this attentional process. Specifically, Mackintosh's (1975) model would predict high attention across the first few trials of overshadowing to both elements of the compound, with a selective decrease to the visual element of the compound. This is despite a relative increase in associative strength attributed to the visual cue from the start of conditioning. Overshadowing of one element of the compound is not predicted by models that utilize a summed-error term (Rescorla and Wagner, 1972; Sutton and Barto, 1981). Rather, models using a summed-error term would predict mutual overshadowing to both elements of the compound. That is, both the salient auditory and less salient visual cue will accrue less associative strength than they would if conditioned individually by virtue of sharing the learning supported by the reward (though the degree to which this occurs is dependent on intrinsic salience). Further, these models are not attentional in nature and would therefore not predict that the signal to either element of the compound would decrease across learning. Thus a search for a Mackintosh (1975) neural signal would have to take into account the complexities of the model rather than just looking for an increase in activity toward predictive cues.

Unifying Models of Attention: Esber and Haselgrove (2011)

So far we have reviewed evidence for each attentional process (Mackintosh, 1975; Pearce and Hall, 1980) as potentially independent yet interactive processes, in line with several hybrid or two-process models of attention (LePelley and McLaren, 2004; Pearce and Mackintosh, 2010). However, another model attempts to reconcile these processes into one mechanism in which attention is directed by *both* predictiveness and uncertainty (Esber and Haselgrove, 2011). Unlike attentional models where the size of the prediction error regulates the amount of attention paid to a cue (Mackintosh, 1975; Pearce and Hall, 1980), the Esber and Haselgrove (2011) model assumes that *acquired salience* of a cue will change with how well it predicts an outcome. At first glance, this sounds similar to Mackintosh's (1975) model of attention. Humans and animals attend to good predictors of reward. However, the Esber and Haselgrove (2011) model also predicts that the omission of an expected reward can function as an effective reinforcer. This is because the frustration caused by omission of an expected

reward is also a motivationally-potent event. Thus, a cue that probabilistically predicts both delivery and omission of expected reward will have increased acquired salience relative to a cue that consistently predicts reward or omission alone, as the former now becomes predictive of two outcomes. Thus, this theory can account for evidence suggesting that humans and animals attend toward good predictors of an outcome (Mackintosh, 1975) while also maintaining attention toward cues which are uncertain predictors of an outcome (Pearce and Hall, 1980).

The critical assumption here is that a cue that is partially reinforced will acquire higher salience relative to a cue that is consistently rewarded (Esber and Haselgrove, 2011). Recently, evidence has emerged showing that some neurons in the orbital frontal cortex (OFC) show such a pattern of responding in anticipation of reward following cue presentation (Ogawa et al., 2013). Most notably, in this study, rats were presented with four odor cues. Two cues consistently predicted reward (100%) or no reward (0%), and two cues inconsistently predicted reward (67%, 33%). Here, around half of the reward-anticipatory neurons in OFC exhibited their highest responding when cues inconsistently predicted reward (67%, 33%). However, critically, these neurons also showed higher firing to certain reward (100%) than certain non-reward (0%), which was near baseline. This pattern- baseline firing in anticipation of non-reward, and increased firing in anticipation of certain reward, and still higher firing to uncertain reward- was perfectly in line with the predictions of the Esber and Haselgrove (2011) model. Future research should explore whether the attentional signal described by Esber and Haselgrove (2011) is pervasive across other systems implicated in attention which may help to reconcile the apparent contradiction in the associative world without appealing to a two-process model. If this is not the case, there are connections between the PL, OFC, and BLA (McDonald, 1987, 1991; Vázquez-Borsetti et al., 2009) that may allow integration of multiple competing processes (see **Figure 2**).

MORE COMPLEX ASSOCIATIVE MODELS

The research above describes how prediction errors may regulate both the rate and amount of learning attributed to a reward-predictive cue across several dissociable circuits. But this is only half the story. Our experience with cues in the environment is often more complex than a discrete cue predicting a rewarding outcome. For one, our experiences are often different depending on context. Consider a veteran coming back from war. During their time at war, they probably formed a strong association between loud noises and negative consequences. However, when the veteran returns home it is far more likely the case that a loud noise signals something innocuous like a slamming door or misfiring engine. It is important in these circumstances that an individual has learned (and can recall) context-specific associations, and does not generalize negative experiences into neutral contexts (Rougemont-Bücking et al., 2011; VanElzakker et al., 2014; Sharpe et al., 2015).

Interestingly, dopamine neurons in the VTA can exhibit context-specific prediction errors that reflect context-specific learning (Nakahara et al., 2004; Kobayashi and Schultz, 2014). For example, Nakahara et al. (2004) trained monkeys to expect reward when presented with a visual cue. Here, one group of monkeys experienced one set of contingencies (a context-independent task), and another group were given another set of contingencies (the context-dependent task). In the 'context-independent' version of the task, the cues were presented with reward 50% of the time, where reward was delivered according to a random distribution. In the 'context-dependent' version of the task, the cues were also reinforced 50% of the time, however, the rate of reinforcement changed depending on the previous run of reinforcement. Here, if monkeys had experienced a long run of non-reinforcement across six trials, they were guaranteed reward on the next trial. So unlike monkeys in the context-independent task, monkeys in the context-dependent task should be able to learn when to expect a rewarded trial. If prediction errors can encode context-dependent information then dopamine activity on the guaranteed rewarded trial after a run of six loses should be minimal, despite the trial constituting an increase in the magnitude of reward that would usually elicit a large prediction error. Sure enough, with extended training prediction errors adjusted to the contextual rule and were modified depending on the prior history of reward. That is, prediction-error signaling was low on trials where monkeys anticipated reward after a long run of unrewarded trials but high when unexpected reward was given before this run of six loses was over. This demonstrates that VTA dopamine prediction-error signals are capable of reflecting information garnered from complex scenarios (Bromberg-Martin et al., 2010a; Takahashi et al., 2011). Since then, it has also been demonstrated that prediction errors can also be modulated by visual background cues (Kobayashi and Schultz, 2014), showing that prediction errors can adjust to both implicit and explicit contextual cues.

Such a finding is compatible with Sutton and Barto's (1981) model-free reinforcement algorithm. This is because this theoretical account relies on the concept of *state*. Here, state is defined as any array of salient observations, either explicit or implicit, that is associated with a particular prediction about the value of upcoming rewards. Hence, during conditioning when a subject experiences presentation of a cue which has been established as predictive of reward, the cue state accrues the value inherent in the reward. Thus, delivery of the reward at the end of cue presentation will not be surprising and a prediction error will not be signaled. Further, the concept of state need not be defined only by reference to the temporally predictive cue. Rather, it can encompass many attributes of the trial. For example, it could include information about how long it has been since reinforcement or other sensory cues (e.g., contextual cues) available on that trial (Nakahara et al., 2004; Redish et al., 2007; Gershman et al., 2010; Nakahara and Hikosaka, 2012; Nakahara, 2014), basically anything that has been directly experienced as associated with reward in the past. Thus, the finding that VTA dopamine prediction-error signals adjust with either implicit or explicit contextual cues can be easily explained within the traditional view that the dopamine error system emits a signal

synonymous with that predicted by model-free algorithms such as that described in Sutton and Barto (1981). This is because different expected values can be assigned to a particular state that are capable of containing information beyond the discrete cue that predicts reward (Bromberg-Martin et al., 2010c; Hong and Hikosaka, 2011; Aitken et al., 2016; Cone et al., 2016).

Not only are dopamine prediction errors capable of reflecting state-specific associations, dopamine prediction errors are also theorized to contribute to the creation of new states which allow for the development of state-specific associations (Gershman et al., 2010; Gershman et al., 2013). Specifically, it is thought that persistently large prediction-error signals may serve as a segmentation signal that alerts the individual to a new state of the world and to form a state-specific association. Take for example, the context-specificity of extinction learning. If a predictive cue is suddenly presented without its predicted outcome, humans and other animals do not unlearn the original cue-outcome association. Rather, they will attribute the change in contingency to any perceived change in the experimental circumstance (Bouton, 2004). Thus, responding to the predictive cue will re-emerge when the experimental circumstance no longer reflects that present in extinction (e.g., the passage of time or a physical change in context; Bouton, 2004). According to Gershman et al. (2010), the large prediction errors present at the beginning of extinction leads an individual to infer a new state and form a context-dependent association specific to the extinction context. In line with this theory, Gershman et al. (2013) have shown that using a gradual extinction procedure, where prediction errors during extinction were reduced by sporadically presenting reinforced trials, reduced the recovery of responding to the predictive cue following the passage of time. This is consistent with an idea that experimentally manipulating the degree of prediction error during extinction reduced the likelihood that a subject will infer a new state and form a context-specific association.

Of course, learning also often extends beyond a reaction to explicit and implicit sensory cues. Humans and other animals are capable of constructing rich associative models of the world which can be flexibly utilized in the absence of direct experience. In such models, a behavioral choice is often made by simulating all possible consequences and selecting the response that is associated with the outcome that is most favorable to the participant. The construction of such models is typically referred to as 'model-based' learning and contains information about value as well as the identity of cues, responses, and rewards. Such learning is typically considered to be independent of a dopaminergic prediction-error system under current interpretations of these signals (Schultz, 1997, 2002, 2007). However, recently research has begun to emerge which suggests that dopaminergic prediction errors may contain model-based information (Daw et al., 2011; Sadacca et al., 2016). For example, dopaminergic prediction errors are influenced by OFC activity, known to be involved in model based behaviors (Takahashi et al., 2011). Further, Daw et al. (2011) recently found evidence for information consistent with a model-based account of behavior in the ventral striatum, traditionally thought to receive a model-free prediction error from VTA dopamine neurons (Suaud-Chagny et al., 1992; Day et al., 2007). Here, they tested human participants on a two-stage decision task. In the first stage, subjects are presented with two pictorial cues. A choice of one cue would lead to a second stage where another set of cues (set 1) are presented the majority of the time, where the choice of the other would lead to a different set of cues (set 2) being presented most of the time. In this second stage, choice of one of the pictorial cues in the two different sets leads to either low or high monetary reward. On rare transitions, the first-stage choice of the set 1 cues would lead to the set 2 of pictorial cues that it is not usually associated with that first-stage choice. The reasoning here is that if the rare transition to the set 2 cues ended up with a choice that lead to an upshift in monetary reinforcement, a model-based agent would select the choice in the first-stage that most likely produces the set 2 cues. That is, they would actually produce a different response from the last reinforced response as it is more often that the alternate choice led to presentation of the set 2 cues. However, a 'model-free' agent, would make the same choice as the last trial. This is because the response on the last trial has just been reinforced and value of that action updated. In line with a model-based account of this behavior, when participants had been reinforced after the rare transition, they choose the different response on first-stage of the next trial that was likely to lead to the pictorial cues that signals greater reinforcement. Further, the Blood Oxygenation Level Dependent (BOLD) activity of this model-based choice were specifically found in ventral striatum, where activity tracked individual differences in degree of model-based behavior. This challenges the traditional assumption that such activity reflects a model-free error signal from VTA dopamine, suggesting this signal facilitates the use of more complex choice behavior that requires an associative structure of the task.

In further support of this notion, Sadacca et al. (2016) have recently found direct evidence that VTA dopamine phasic signals in the rodent encodes model-based information. Using a sensory-preconditioning task, Sadacca et al. (2016) found that VTA dopamine neurons emit their traditional phasic signal toward a cue that has not been directly paired with reward but, rather, has come to predict reward via its associative relationship with another reward-paired cue. Sensory preconditioning involves first pairing two neutral cues as a serial compound in the absence of any reward. Following this preconditioning phase, one of these cues is then paired directly with reward during conditioning. As a consequence of this training, both the reward-paired and neutral cue will now elicit the expectation of reward. Thus, the cue not directly paired with reward also acquires an ability to predict reward via its prior association with the to-be-conditioned cue. Such a prediction is model-based as updating learning in the absence of direct experience requires the existence of a mental map of relationships between cues that can be flexibly adapted to incorporate the new information. Interestingly, Sadacca et al. (2016) found that VTA dopamine neurons responded to both the cue directly paired with reward and the neutral cue that came to predict reward by virtue its associative link with the reward-paired cue in the preconditioning phase. This data clearly demonstrates that VTA dopamine neurons encode associations that reflect model-based inference not based on

direct experience. Thus emerging evidence from both the human and rodent literature has begun to suggest that the dopaminergic prediction-error system contains information that goes above and beyond that appropriately described as a model-free value signal described in Sutton and Barto (1981).

OUTSTANDING QUESTIONS

How Does the Diversity of VTA Dopamine Neurons and Their Projection Targets Lend to Our Understanding of How Associative Learning Systems Interact?

A growing interest in the field is the investigation of the heterogeneity of dopaminergic neurons in the VTA and the diversity of their neurons targets (Lammel et al., 2011; Parker et al., 2016; Morales and Margolis, 2017). For example, studies have identified that distinct populations of dopamine neurons in the VTA that show preferential increases in firing to either rewarding and aversive outcomes and the cues which predict their occurrence (Matsumoto and Hikosaka, 2009; Bromberg-Martin et al., 2010b). In parallel, research has shown that distinct populations of VTA dopamine neurons receive input from the laterodorsal tegmentum (LDT) and lateral habenula (LHb), argued to underlie these appetitive and aversive responses, respectively (Lammel et al., 2012). These inputs from LDT and LHb synapse preferentially on VTA dopamine neurons projecting to nucleus accumbens (NAc) and PFC, respectively. These studies are a few of a host of studies which are beginning to identify disparate populations of VTA dopamine neurons that appear to show distinct and complex interactions with wider neuronal systems where they contribute to behavior in diverse ways (Matsumoto and Hikosaka, 2007; Jhou et al., 2009; Lammel et al., 2011; Eban-Rothschild et al., 2016; Parker et al., 2016). Additional complexity of the VTA dopamine system comes from recent evidence which suggests that VTA dopmaine neurons also release other neurotransmitters such as glutamate and GABA. Thus this emerging research begins to paint a complex picture of how VTA dopamine neurons may contribute to learning and behavior which may continue to challenge a perception of the prediction error as a cached-value signal. The continuation of such research will undoubtedly shed light on the ways in which VTA dopamine prediction-error signaling contributes to attentional and model-based learning described in this review.

What about Non-dopaminergic VTA Neurons?

The prediction-error signal to the reward wanes across successive cue-reward parings as the cue comes to reliably predict the reward. However, with this decrease in signal at the time of the reward, we also see an increase of dopamine signaling to the reward-predictive cue. This phasic response to the cue is thought to reflect the cached value inherent in the reward it predicts. It has been suggested that this reduction in the neural response at the time of reward, as a result of expectation elicited by a cue, may arise from inhibition of dopamine neurons that is initiated after cue offset and persists during reward (see **Figure 3**). GABAergic neurons in the VTA are one possible candidate proposed to provide this inhibitory signal. Recently, Cohen et al. (2012)

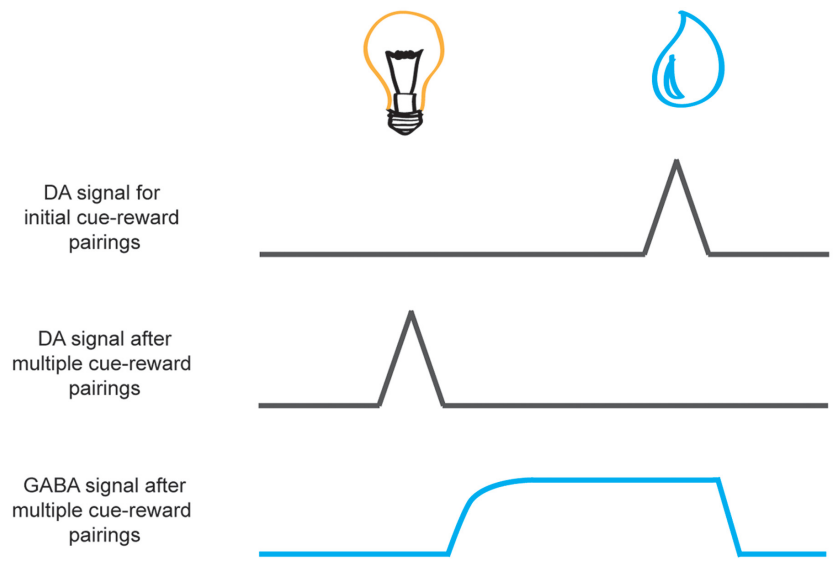


FIGURE 3 | Schematic of dopamine and GABA reward prediction-error activity during learning. Neural activity is aligned to cue presentation (e.g., light, on the left) and reward presentation (e.g., a drop of juice, on the right). While phasic activity of dopamine neurons (black lines) are elicited by unexpected reward delivery upon initial cue-reward pairings (top) with repeated cue-reward pairings the signal at the time of reward receipt wanes as the reward becomes predicted by the cue (middle). This transition occurs gradually over successive trials in accordance with traditional learning models of prediction error (Rescorla and Wagner, 1972; Sutton and Barto, 1981). It is speculated that this reduction in the dopamine signal to the reward may result from inhibition of dopamine neurons by GABAergic neurons in the VTA (bottom, blue line) that is initiated after cue offset and persists during reward delivery (Houk et al., 1995; Cohen et al., 2012).

recorded GABAergic neurons in animals well trained on a simple cue-reward procedure where different odor cues predicted either big reward, small reward, nothing, or punishment. Cohen et al. (2012) found that dopaminergic neurons responded to cues in a manner consistent with the quantitative value it predicted. However, while GABAergic neurons were excited by predictive cues, they exhibited sustained activity across the delay between the cue and the expected reward (see **Figure 3**). The authors concluded that this signal from GABAergic neurons counteracts the excitatory drive of dopaminergic neurons when a reward has been predicted to ensure that a prediction-error signal is not elicited when an expected reward is delivered. Thus, these GABAergic neurons may contribute to the development of a reduction in the dopaminergic response during reward receipt as the cue comes to predict the reward. Future studies may continue to investigate the causal role of these neurons in learning and to determine which inputs from other regions provide the expectancy signal to allow GABAergic neurons to modulate dopaminergic prediction-error signals in the VTA.

Is Learning Always Distributed in Accordance with a Summed-Error Term?

As it stands we have argued that dopamine signaling in the VTA can support learning in a manner that is consistent with multiple theories of associative learning. In doing so, we have predominantly focused on how VTA dopamine may relay a summed-error term to facilitate cue processing in other brain regions (Rescorla and Wagner, 1972; Sutton and Barto, 1998). However, empirical data has shown that learning can also be governed by an individual-error term; as such learning on any one trial need not be equally distributed across cues present on a trial even if they are of equal salience (Le Pelley and McLaren, 2001; LePelley and McLaren, 2004; Leung and Westbrook, 2008). One of the most convincing findings in favor of the presence of individual-error terms comes from studies of causal learning in humans (LePelley and McLaren, 2004). Specifically, Le Pelley and McLaren (2001) looked at the distribution of associative change between the elements of a compound composed of an excitatory cue and an inhibitory cue. In contrast to the predictions made by models comprising a summed-error term (Rescorla and Wagner, 1972; Sutton and Barto, 1981), they found that learning was not distributed equally across the elements of the compound. When the compound was reinforced, the excitatory cue underwent greater change, however, when the compound was not reinforced the inhibitory cue underwent greater change. These data cannot be accounted for by a summed-error term (nor a differential degree of attention directed toward one of the cues). Rather, these data suggest that an individual-error term must be at least capable of contributing to associative change in some settings. As a consequence of such evidence, more recent developments in models of associative learning have taken into account the need for individual-error terms (Mackintosh, 1975; Pearce and Hall, 1980; Rescorla, 2000; LePelley and McLaren, 2004; Pearce and Mackintosh, 2010; Le Pelley et al., 2012). While there has been little investigation into the neural mechanism underlying individual-error terms, it would be of interest to identify whether midbrain dopamine signals may also reflect an individual-error term to contribute to associative change under these circumstances.

How Might We Reconcile Evidence for Model-Based Learning in the VTA within the Current Framework?

Of course, a discussion of how VTA dopamine signaling impacts other structures to produce many forms of learning driven by error correction is a one-sided view. VTA dopamine neurons not only project out to a rich neural circuit, they receive dense reciprocal projections from these regions (Carr and Sesack, 2000; Vázquez-Borsetti et al., 2009, see **Figure 2**). Taking the broader circuitry into account, perhaps areas known to be involved in model-based reasoning inform VTA dopamine phasic signals of learning outcomes garnered from more flexible mental representations developed in the absence of direct experience. Thus, this information could be relayed in a top-down manner to VTA to modulate these phasic signals according to this word view (Daw et al., 2011; Takahashi et al., 2011, 2016; O'Doherty et al., 2017). However, it is also possible that VTA dopamine signals are causally involved in promoting the development of the associations which underlie the development of flexible mental maps which facilitate model-based inference. That is, these signals may provide more complex associative information about relationships between cues and outcome that facilitate model-based behaviors. While we have begun to scratch the surface of how dopamine signaling may influence model-based mechanisms, we need to start causally testing predictions of dopamine functioning beyond that envisioned by Sutton and Barto's (1981) model-free reinforcement learning algorithm to truly understand all the weird and wonderful ways that phasic VTA dopamine supports associative learning.

AUTHOR CONTRIBUTIONS

HN and MS wrote major parts of the article. All other authors critically reviewed and edited the article. The review was written based on the expertise of the authors, who have sourced the article on PubMed and Google Scholar.

ACKNOWLEDGMENT

The work was supported by the Intramural Research Program of the National Institute on Drug Abuse.

REFERENCES

Aitken, T. J., Greenfield, V. Y., and Wassum, K. M. (2016). Nucleus accumbens core dopamine signaling tracks the need-based motivational value of food-paired cues. *J. Neurochem.* 136, 1026–1036. doi: 10.1111/jnc.13494

Arcediano, F., Escobar, M., and Miller, R. R. (2004). Is stimulus competition an acquisition deficit or a performance deficit? *Psychon. Bull. Rev.* 11, 1105–1110. doi: 10.3758/BF03196744

Bayer, H. M., and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47, 129–141. doi: 10.1016/j.neuron.2005.05.020

Belova, M. A., Paton, J. J., Morrison, S. E., and Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron* 55, 970–984. doi: 10.1016/j.neuron.2007.08.004

Belova, M. A., Paton, J. J., and Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *J. Neurosci.* 28, 10023–10030. doi: 10.1523/JNEUROSCI.1400-08.2008

Bentivoglio, M., and Morelli, M. (2005). Chapter I The organization and circuits of mesencephalic dopaminergic neurons and the distribution of dopamine receptors in the brain. *Handb. Chem. Neuroanat.* 21, 1–107. doi: 10.1016/S0924-8196(05)80005-3

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *Eur. J. Neurosci.* 35, 1124–1143. doi: 10.1111/j.1460-9568.2012.07990.x

Beyeler, A., Namburi, P., Glober, G. F., Simonnet, C., Calhoon, G. G., Conyers, G. F., et al. (2016). Divergent routing of positive and negative information from the amygdala during memory retrieval. *Neuron* 90, 348–361. doi: 10.1016/j.neuron.2016.03.004

Birrell, J. M., and Brown, V. J. (2000). Medial frontal cortex mediates perceptual attentional set shifting in the rat. *J. Neurosci.* 20, 4320–4324.

Björklund, A., and Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends Neurosci.* 30, 194–202. doi: 10.1016/j.tins.2007.03.006

Blaisdell, A. P., Denniston, J. C., and Miller, R. R. (1999). Posttraining shifts in the overshadowing stimulus-unconditioned stimulus interval alleviates the overshadowing deficit. *J. Exp. Psychol. Anim. Behav. Process.* 25, 18–27. doi: 10.1037/0097-7403.25.1.18

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learn. Mem.* 11, 485–494. doi: 10.1101/lm.78804

Brischoux, F., Chakraborty, S., Brierley, D. I., and Ungless, M. A. (2009). Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. *Proc. Natl. Acad. Sci. U.S.A.* 106, 4894–4899. doi: 10.1073/pnas.0811507106

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010a). Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. *Neuron* 67, 144–155. doi: 10.1016/j.neuron.2010.06.016

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010b). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron* 68, 815–834. doi: 10.1016/j.neuron.2010.11.022

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010c). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.* 104, 1068–1076. doi: 10.1152/jn.00158.2010

Bush, R. R., and Mosteller, F. (1951). A mathematical model for simple learning. *Psychol. Rev.* 58, 313–323. doi: 10.1037/H0054388

Carr, D. B., and Sesack, S. R. (2000). Projections from the rat prefrontal cortex to the ventral tegmental area: target specificity in the synaptic associations with mesoaccumbens and mesocortical neurons. *J. Neurosci.* 20, 3864–3873.

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H. J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nat. Neurosci.* 19, 111–116. doi: 10.1038/nn.4191

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 482, 85–88. doi: 10.1038/nature10754

Cone, J. J., Fortin, S. M., McHenry, J. A., Stuber, G. D., McCutcheon, J. E., and Roitman, M. F. (2016). Physiological state gates acquisition and expression of mesolimbic reward prediction signals. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1943–1948. doi: 10.1073/pnas.1519643113

Corlett, P. R., Murray, G. K., Honey, G. D., Aitken, M. R., Shanks, D. R., Robbins, T. W., et al. (2007). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain* 130, 2387–2400. doi: 10.1093/brain/awm173

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215. doi: 10.1016/j.neuron.2011.02.027

Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* 10, 1020–1028. doi: 10.1038/nn1923

Dayan, P., and Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.* 14, 473–492. doi: 10.3758/s13415-014-0277-8

Dickinson, A., and Balleine, B. (2002). "The role of learning in the operation of motivational systems," in *Steven's Handbook of Experimental Psychology*, 3rd Edn, eds H. Pashler and R. Gallistel (New York, NY: John Wiley & Sons, Inc.), 497–533.

Eban-Rothschild, A., Rothschild, G., Giardino, W. J., Jones, J. R., and de Lecea, L. (2016). VTA dopaminergic neurons regulate ethologically relevant sleep-wake behaviors. *Nat. Neurosci.* 19, 1356–1366. doi: 10.1038/nn.4377

Esber, G. R., and Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proc. Biol. Sci.* 278, 2553–2561. doi: 10.1098/rspb.2011.0836

Esber, G. R., and Holland, P. C. (2014). The basolateral amygdala is necessary for negative prediction errors to enhance cue salience, but not to produce conditioned inhibition. *Eur. J. Neurosci.* 40, 3328–3337. doi: 10.1111/ejn.12695

Esber, G. R., Roesch, M. R., Bali, S., Trageser, J., Bissonette, G. B., Puche, A. C., et al. (2012). Attention-related Pearce-Kaye-Hall signals in basolateral amygdala require the midbrain dopaminergic system. *Biol. Psychiatry* 72, 1012–1019. doi: 10.1016/j.biopsych.2012.05.023

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* 19, 479–486. doi: 10.1038/nn.4239

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898–1902. doi: 10.1126/science.1077349

Gershman, S. J., Blei, D. M., and Niv, Y. (2010). Context, learning, and extinction. *Psychol. Rev.* 117, 197–209. doi: 10.1037/a0017808

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M. H., and Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Front. Behav. Neurosci.* 7:164. doi: 10.3389/fnbeh.2013.00164

Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron* 76, 281–295. doi: 10.1016/j.neuron.2012.09.034

Gottlieb, J., Hayhoe, M., Hikosaka, O., and Rangel, A. (2014). Attention, reward, and information seeking. *J. Neurosci.* 34, 15497–15504. doi: 10.1523/JNEUROSCI.3270-14.2014

Griffiths, O., Johnson, A. M., and Mitchell, C. J. (2011). Negative transfer in human associative learning. *Psychol. Sci.* 22, 1198–1204. doi: 10.1177/0956797611419305

Gruber, A. J., Calhoon, G. G., Shusterman, I., Schoenbaum, G., Roesch, M. R., and O'Donnell, P. (2010). More is less: a disinhibited prefrontal cortex impairs cognitive flexibility. *J. Neurosci.* 30, 17102–17110. doi: 10.1523/JNEUROSCI.4623-10.2010

Hall, G., and Pearce, J. M. (1979). Latent inhibition of a CS during CS–US pairings. *J. Exp. Psychol. Anim. Behav. Process.* 5, 31–42. doi: 10.1037/0097-7403.5.1.31

Hare, T. A., Malmaud, J., and Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. *J. Neurosci.* 31, 11077–11087. doi: 10.1523/JNEUROSCI.6383-10.2011

Herry, C., Bach, D. R., Esposito, F., Di Salle, F., Perrig, W. J., Scheffler, K., et al. (2007). Processing of temporal unpredictability in human and animal amygdala. *J. Neurosci.* 27, 5958–5966. doi: 10.1523/JNEUROSCI.5218-06.2007

Hickey, C., Kaiser, D., and Peelen, M. V. (2015). Reward guides attention to object categories in real-world scenes. *J. Exp. Psychol. Gen.* 144, 264–273. doi: 10.1037/a0038627

Hickey, C., McDonald, J. J., and Theeuwes, J. (2006). Electrophysiological evidence of the capture of visual attention. *J. Cogn. Neurosci.* 18, 604–613. doi: 10.1162/jocn.2006.18.4.604

Hickey, C., Olivers, C., Meeter, M., and Theeuwes, J. (2011). Feature priming and the capture of visual attention: linking two ambiguity resolution hypotheses. *Brain Res.* 1370, 175–184. doi: 10.1016/j.brainres.2010.11.025

Hickey, C., and Theeuwes, J. (2011). Context and competition in the capture of visual attention. *Atten. Percept. Psychophys.* 73, 2053–2064. doi: 10.3758/s13414-011-0168-9

Holland, P. C., Chik, Y., and Zhang, Q. (2001). Inhibitory learning tests of conditioned stimulus associability in rats with lesions of the amygdala central nucleus. *Behav. Neurosci.* 115, 1154–1158. doi: 10.1037/0735-7044.115.5.1154

Holland, P. C., and Gallagher, M. (1993). Amygdala central nucleus lesions disrupt increments, but not decrements, in conditioned stimulus processing. *Behav. Neurosci.* 107, 246–253. doi: 10.1037/0735-7044.107.2.246

Holland, P. C., and Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends Cogn. Sci.* 3, 65–73. doi: 10.1016/S1364-6613(98)01271-6

Holland, P. C., and Schiffino, F. L. (2016). Mini-review: prediction errors, attention and associative learning. *Neurobiol. Learn. Mem.* 131, 207–215. doi: 10.1016/j.nlm.2016.02.014

Hollerman, J. R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* 1, 304–309. doi: 10.1038/1124

Hong, S., and Hikosaka, O. (2011). Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning. *Front. Behav. Neurosci.* 5:15. doi: 10.3389/fnbeh.2011.00015

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). "A model of how the basal ganglia generates and uses neural signals that predict reinforcement," in *Models of Information Processing in the Basal Ganglia*, eds J. C. Houk, J. Davis, and D. Beiser (Cambridge, MA: MIT Press), 249–270.

Iordanova, M. D. (2009). Dopaminergic modulation of appetitive and aversive predictive learning. *Rev. Neurosci.* 20, 383–404. doi: 10.1515/REVNEURO.2009.20.5-6.383

Jhou, T. C., Fields, H. L., Baxter, M. G., Saper, C. B., and Holland, P. C. (2009). The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. *Neuron* 61, 786–800. doi: 10.1016/j.neuron.2009.02.001

Jovancevic, J., Sullivan, B., and Hayhoe, M. (2006). Control of attention and gaze in complex environments. *J. Vis.* 6, 1431–1450. doi: 10.1167/6.12.9

Kamin, L. J. (1968). Attention-like" processes in classical conditioning. *Paper Presented at the Miami Symposium on the Prediction of Behavior: Aversive Stimulation* (Miami, FL: University of Miami Press).

Keiflin, R., and Janak, P. H. (2015). Dopamine prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron* 88, 247–263. doi: 10.1016/j.neuron.2015.08.037

Kobayashi, S., and Schultz, W. (2014). Reward contexts extend dopamine signals to unrewarded stimuli. *Curr. Biol.* 24, 56–62. doi: 10.1016/j.cub.2013.10.061

Lak, A., Stauffer, W. R., and Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2343–2348. doi: 10.1073/pnas.1321596111

Lammel, S., Hetzel, A., Hackel, O., Jones, I., Liss, B., and Roeper, J. (2008). Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron* 57, 760–773. doi: 10.1016/j.neuron.2008.01.022

Lammel, S., Ion, D. I., Roeper, J., and Malenka, R. C. (2011). Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron* 70, 855–862. doi: 10.1016/j.neuron.2011.03.025

Lammel, S., Lim, B. K., and Malenka, R. C. (2014). Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* 76(Pt B), 351–359. doi: 10.1016/j.neuropharm.2013.03.019

Lammel, S., Lim, B. K., Ran, C., Huang, K. W., Betley, M. J., Tye, K. M., et al. (2012). Input-specific control of reward and aversion in the ventral tegmental area. *Nature* 491, 212–217. doi: 10.1038/nature11527

Le Pelley, M., Beesley, T., and Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *J. Exp. Psychol.* 37, 220–229. doi: 10.1037/a0021384

Le Pelley, M., and McLaren, I. (2001). Retrospective revaluation in humans: learning or memory? *Q. J. Exp. Psychol.* 54, 311–352. doi: 10.1080/02724990143000072

Le Pelley, M. E., Haselgrove, M., and Esber, G. R. (2012). Modeling attention in associative learning: two processes or one? *Learn. Behav.* 40, 292–304. doi: 10.3758/s13420-012-0084-4

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., and Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychol. Bull.* 142, 1111–1140. doi: 10.1037/bul0000064

Le Pelley, M. E., Vadillo, M., and Luque, D. (2013). Learned predictiveness influences rapid attentional capture: evidence from the dot probe task. *J. Exp. Psychol.* 39, 1888–1900. doi: 10.1037/a0033700

Lee, H. J., Youn, J. M., O, M. J., Gallagher, M., and Holland, P. C. (2006). Role of substantia nigra-amygdala connections in surprise-induced enhancement of attention. *J. Neurosci.* 26, 6077–6081. doi: 10.1523/JNEUROSCI.1316-06.2006

LePelley, M. E., and McLaren, I. P. (2004). Associative history affects the associative change undergone by both presented and absent cues in human causal learning. *J. Exp. Psychol. Anim. Behav. Process.* 30, 67–73. doi: 10.1037/0097-7403.30.1.67

Leung, H. T., and Westbrook, R. F. (2008). Spontaneous recovery of extinguished fear responses deepens their extinction: a role for error-correction mechanisms. *J. Exp. Psychol.* 34, 461–474. doi: 10.1037/0097-7403.34.4.461

Lewis, B. L., and O'Donnell, P. (2000). Ventral tegmental area afferents to the prefrontal cortex maintain membrane potential 'up'states in pyramidal neurons via D1 dopamine receptors. *Cereb. Cortex* 10, 1168–1175. doi: 10.1093/cercor/10.12.1168

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., and Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nat. Neurosci.* 14, 1250–1252. doi: 10.1038/nn.2904

Lim, S. L., O'Doherty, J. P., and Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *J. Neurosci.* 31, 13214–13223. doi: 10.1523/JNEUROSCI.1246-11.2011

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* 67, 145–163.

Mackintosh, N. J. (1965). Selective attention in animal discrimination learning. *Psychol. Bull.* 64, 124. doi: 10.1037/h0022347

Mackintosh, N. J. (1973). "Stimulus selection: learning to ignore stimuli that predict no change in reinforcement," in *Constraints on Learning*, eds R. A. Hinde and J. Stevenson-Hinde (New York, NY: Academic Press), 75–96.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychol. Rev.* 82, 276–298. doi: 10.1037/h0076778

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Anim. Learn. Behav.* 4, 186–192. doi: 10.3758/bf03214033

Margolis, E. B., Lock, H., Hjelmstad, G. O., and Fields, H. L. (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *J. Physiol.* 577(Pt 3), 907–924. doi: 10.1113/jphysiol.2006.117069

Matsumoto, M., and Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* 447, 1111–1115. doi: 10.1038/nature05860

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* 459, 837–841. doi: 10.1038/nature08028

McDonald, A. J. (1987). Organization of amygdaloid projections to the mediodorsal thalamus and prefrontal cortex: a fluorescence retrograde transport study in the rat. *J. Comp. Neurol.* 262, 46–58. doi: 10.1002/cne.902620105

McDonald, A. J. (1991). Organization of amygdaloid projections to the prefrontal cortex and associated striatum in the rat. *Neuroscience* 44, 1–14. doi: 10.1016/0306-4522(91)90247-L

McLaren, I., and Mackintosh, N. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Anim. Learn. Behav.* 30, 177–200. doi: 10.3758/BF03192828

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Miller, R. R., and Matzel, L. D. (1988). The comparator hypothesis: a response rule for the expression of associations. *Psychol. Learn. Motiv.* 22, 51–92. doi: 10.1016/S0079-7421(08)60038-9

Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.* 72, 1024–1027.

Mirenowicz, J., and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* 379, 449–451. doi: 10.1038/379449a0

Morales, M., and Margolis, E. B. (2017). Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nat. Rev. Neurosci.* 18, 73–85. doi: 10.1038/nrn.2016.165

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* 9, 1057–1063. doi: 10.1038/nn1743

Nakahara, H. (2014). Multiplexing signals in reinforcement learning with internal models and dopamine. *Curr. Opin. Neurobiol.* 25, 123–129. doi: 10.1016/j.conb.2014.01.001

Nakahara, H., and Hikosaka, O. (2012). Learning to represent reward structure: a key to adapting to complex environments. *Neurosci. Res.* 74, 177–183. doi: 10.1016/j.neures.2012.09.007

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron* 41, 269–280. doi: 10.1016/S0896-6273(03)00869-9

Niki, H., and Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Res.* 171, 213–224. doi: 10.1016/0006-8993(79)90328-7

Niv, Y., and Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends Cogn. Sci.* 12, 265–272. doi: 10.1016/j.tics.2008.03.006

O'Doherty, J. P., Cockburn, J., and Pauli, W. M. (2017). Learning, reward, and decision making. *Annu. Rev. Psychol.* 68, 73–100. doi: 10.1146/annurev-psych-010416-044216

Ogawa, M., van der Meer, M. A., Esber, G. R., Cerri, D. H., Stalnaker, T. A., and Schoenbaum, G. (2013). Risk-responsive orbitofrontal neurons track acquired salience. *Neuron* 77, 251–258. doi: 10.1016/j.neuron.2012.11.006

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., et al. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nat. Neurosci.* 19, 845–854. doi: 10.1038/nn.4287

Pearce, J. M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87, 532–552. doi: 10.1037/0033-295X.87.6.532

Pearce, J. M., and Mackintosh, N. J. (2010). "Two theories of attention: a review and a possible integration," in *Attention and Associative Learning: From Brain to Behaviour*, eds C. J. Mitchell and M. E. Le Pelley (Oxford: Oxford University Press), 11–39.

Pitkanen, A., Jolkkonen, E., and Kemppainen, S. (2000). Anatomic heterogeneity of the rat amygdaloid complex. *Folia Morphol.* 59, 1–24.

Redish, A. D., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol. Rev.* 114, 784–805. doi: 10.1037/0033-295X.114.3.784

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *J. Exp. Psychol. Anim. Behav. Process.* 26, 428–438. doi: 10.1037/0097-7403.26.4.428

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, Vol. 2, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton Century Crofts), 64–99. doi: 10.1037/a0030892

Roesch, M. R., Calu, D. J., Esber, G. R., and Schoenbaum, G. (2010). Neural correlates of variations in event processing during learning in basolateral amygdala. *J. Neurosci.* 30, 2464–2471. doi: 10.1523/JNEUROSCI.5781-09.2010

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* 10, 1615–1624. doi: 10.1038/nn2013

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., and Schoenbaum, G. (2012). Surprise! neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *Eur. J. Neurosci.* 35, 1190–1200. doi: 10.1111/j.1460-9568.2011.07986.x

Rougemont-Bücking, A., Linnman, C., Zeffiro, T. A., Zeidan, M. A., Lebron-Milad, K., Rodriguez-Romaguera, J., et al. (2011). Altered processing of contextual information during fear extinction in PTSD: an fMRI study. *CNS Neurosci. Ther.* 17, 227–236. doi: 10.1111/j.1755-5949.2010.00152.x

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., and Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron* 70, 1054–1069. doi: 10.1016/j.neuron.2011.05.014

Sadacca, B. F., Jones, J. L., and Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife* 5:e13665. doi: 10.7554/eLife.13665

Schultz, W. (1997). Dopamine neurons and their role in reward mechanisms. *Curr. Opin. Neurobiol.* 7, 191–197. doi: 10.1016/S0959-4388(97)80007-4

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron* 36, 241–263. doi: 10.1016/S0896-6273(02)00967-4

Schultz, W. (2007). Behavioral dopamine signals. *Trends Neurosci.* 30, 203–210. doi: 10.1016/j.tins.2007.03.007

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nat. Rev. Neurosci.* 17, 183–195. doi: 10.1038/nrn.2015.26

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Sharpe, M. J., and Killcross, S. (2014). The prelimbic cortex contributes to the down-regulation of attention toward redundant cues. *Cereb. Cortex* 24, 1066–1074. doi: 10.1093/cercor/bhs393

Sharpe, M. J., and Killcross, S. (2015). The prelimbic cortex directs attention toward predictive cues during fear learning. *Learn. Mem.* 22, 289–293. doi: 10.1101/lm.038273.115

Sharpe, M. J., Wikenheiser, A. M., Niv, Y., and Schoenbaum, G. (2015). The state of the orbitofrontal cortex. *Neuron* 88, 1075–1077. doi: 10.1016/j.neuron.2015.12.004

Stauffer, W. R., Lak, A., and Schultz, W. (2014). Dopamine reward prediction error responses reflect marginal utility. *Curr. Biol.* 24, 2491–2500. doi: 10.1016/j.cub.2014.08.064

Stauffer, W. R., Lak, A., Yang, A., Borel, M., Paulsen, O., Boyden, E. S., et al. (2016). Dopamine neuron-specific optogenetic stimulation in rhesus macaques. *Cell* 166, 1564–1571.e6. doi: 10.1016/j.cell.2016.08.024

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* 16, 966–973. doi: 10.1038/nn.3413

Suaud-Chagny, M., Chergui, K., Chouvet, G., and Gonon, F. (1992). Relationship between dopamine release in the rat nucleus accumbens and the discharge activity of dopaminergic neurons during local in vivo application of amino acids in the ventral tegmental area. *Neuroscience* 49, 63–72. doi: 10.1016/0306-4522(92)90076-E

Sutton, R. S., and Barto, A. G. (1981). Toward a modern theory of adaptive networks – expectation and prediction. *Psychol. Rev.* 88, 135–170. doi: 10.1037/0033-295x.88.2.135

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, Vol. 1. Cambridge, MA: MIT Press.

Swanson, L. W. (1982). The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Res. Bull.* 9, 321–353. doi: 10.1016/0361-9230(82)90145-9

Takahashi, Y. K., Langdon, A. J., Niv, Y., and Schoenbaum, G. (2016). Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. *Neuron* 91, 182–193. doi: 10.1016/j.neuron.2016.05.015

Takahashi, Y. K., Roesch, M. R., Wilson, R. C., Toreson, K., O'Donnell, P., Niv, Y., et al. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci.* 14, 1590–1597. doi: 10.1038/nn.2957

Theeuwes, J. (2013). Feature-based attention: it is all bottom-up priming. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368:20130055. doi: 10.1098/rstb.2013.0055

Tobler, P. N., Dickinson, A., and Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *J. Neurosci.* 23, 10402–10410.

Tommasi, G., Fiorio, M., Yelnik, J., Krack, P., Sala, F., Schmitt, E., et al. (2015). Disentangling the role of cortico-basal ganglia loops in top-down and bottom-up visual attention: an investigation of attention deficits in Parkinson disease. *J. Cogn. Neurosci.* 27, 1215–1237. doi: 10.1162/jocn_a_00770

Tsai, H. C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., et al. (2009). Phasic firing in dopaminergic neurons is sufficient

for behavioral conditioning. *Science* 324, 1080–1084. doi: 10.1126/science.1168878

Tye, K. M., Cone, J. J., Schairer, W. W., and Janak, P. H. (2010). Amygdala neural encoding of the absence of reward during extinction. *J. Neurosci.* 30, 116–125. doi: 10.1523/JNEUROSCI.4240-09.2010

Tzschentke, T. M., and Schmidt, W. J. (2000). Functional relationship among medial prefrontal cortex, nucleus accumbens, and ventral tegmental area in locomotion and reward. *Crit. Rev. Neurobiol.* 14, 131–142. doi: 10.1615/CritRevNeurobiol.v14.i2.20

VanElzakker, M. B., Dahlgren, M. K., Davis, F. C., Dubois, S., and Shin, L. M. (2014). From Pavlov to PTSD: the extinction of conditioned fear in rodents, humans, and anxiety disorders. *Neurobiol. Learn. Mem.* 113, 3–18. doi: 10.1016/j.nlm.2013.11.014

Vázquez-Borsetti, P., Cortés, R., and Artigas, F. (2009). Pyramidal neurons in rat prefrontal cortex projecting to ventral tegmental area and dorsal raphe nucleus express 5-HT2A receptors. *Cereb. Cortex* 19, 1678–1686. doi: 10.1093/cercor/bhn204

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412, 43–48. doi: 10.1038/35083500

Wilschut, A., Theeuwes, J., and Olivers, C. N. (2015). Nonspecific competition underlies transient attention. *Psychol. Res.* 79, 844–860. doi: 10.1007/s00426-014-0605-1

Wilson, P. N., Boumphrey, P., and Pearce, J. M. (1992). Restoration of the orienting response to a light by a change in its predictive accuracy. *Q. J. Exp. Psychol.* 44, 17–36.

Witten, I. B., Steinberg, E. E., Lee, S. Y., Davidson, T. J., Zalocusky, K. A., Brodsky, M., et al. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron* 72, 721–733. doi: 10.1016/j.neuron.2011.10.028

Conflict of Interest Statement: 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.

Copyright © 2017 Nasser, Calu, Schoenbaum and Sharpe. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on