frontiers | Frontiers in Education

TYPE Conceptual Analysis
PUBLISHED 27 November 2025
DOI 10.3389/feduc.2025.1677255

OPEN ACCESS

EDITED BY
Leonora Kaldaras,
Texas Tech University, United States

REVIEWED BY
Vanessa Scherman,
International Baccalaureate (IBO), Netherlands
Cesare Aloisi,
Assessment and Qualifications Alliance Manchester Office, United Kingdom

*CORRESPONDENCE
Max van Haastrecht
✉ max.vanhaastrecht@cito.nl

RECEIVED 31 July 2025
REVISED 04 October 2025
ACCEPTED 04 November 2025
PUBLISHED 27 November 2025

CITATION
van Haastrecht M, de Groot L, Jongbloed-Pereboom M, Buytenhuijs F and Kruis J (2025) Artificial intelligence for educational measurement: Where is the value for education? *Front. Educ.* 10:1677255. doi: 10.3389/feduc.2025.1677255

Artificial intelligence for educational measurement: Where is the value for education?

Max van Haastrecht*, Lotte de Groot, Marjolein Jongbloed-Pereboom, Franka Buytenhuijs and Joost Kruis

CitoLab, Stichting Cito, Arnhem, Netherlands

Artificial intelligence (AI) systems are not intrinsically valuable to education, but rather lend their value from contributing to educational goals. However, when it comes to educational measurement, it is often unclear whether and how AI systems help us to achieve our goals. In this paper, we introduce a way of thinking that helps to clarify how the rules and structures governing educational assessments are impacted by AI systems. Based on a conceptual analysis of the literature, we outline three core elements that should be contemplated when integrating AI systems into assessment: the educational measurement context, the prioritization of different facets of assessment validity, and the social contract between student and teacher. We apply our way of thinking to analyze case studies of AI in item construction, assessing written work, and grading assistance. We show how requiring active reflection on educational aims can inform the realization that gains in subsidiary aims such as efficiency do not provide sufficient warrant for making the move toward AI. We hope that this new way of thinking can instigate critical reflection on what we value in education and how assessments can be designed to reflect those values.

KEYWORDS

educational measurement, artificial intelligence, validity, social contract, educational assessment

1 Introduction

Education is at its core a human endeavor, rooted in the interaction between students and teachers. This perhaps explains the uncomfortable feeling many educators have when technological innovations threaten to redefine the human role in education. Nowhere is this feeling more prominent today than with developments in artificial intelligence (AI). There is a consensus that we should only use AI in education if it aids learning and teaching processes (Molenaar, 2022), but different educators have different beliefs about what the right processes are and what constitutes good education (Biesta, 2015). A discussion on whether to use AI for a specific task and context should therefore always be preceded by a discussion of what the people involved in that context value in education.

When we have a conceptualization of what we want to see students achieve and how we want to see them grow, we can start to ask how exactly we can realize this "seeing". The issue at hand is that we cannot directly observe a student's knowledge, skills, and attitudes (KSAs) (Arieli-Attali et al., 2019). Instead, we must uncover these hidden and unknown KSAs as best we can through a process of inquiry. Educational measurement is this process of inquiry (Dewey, 1910; Delandshere, 2002), where the aim is to clarify the state and trajectory of KSAs, enabling alignment with educational goals.

Educational measurement is thus a means to the end of good education, in the same way that AI can be a means to the end of good educational measurement.

It is by no means guaranteed that AI interventions will positively impact educational measurement. Given the challenges surrounding bias, explainability, and reliability associated to current AI systems, we risk losing the trust built up in assessments by implementing AI without proper forethought (Aloisi, 2023). These challenges are especially acute in generative AI systems based on large language models, that are by their nature probabilistic and complex. We should thus ask how, when, and where AI facilitates better education through its use in educational measurement. In this paper, we introduce a way of thinking that helps to understand in which situations AI can add value to education through improved educational measurement, and in which situations AI is more harmful than helpful. This way of thinking is founded on two main premises. Firstly, that technological innovations such as AI are not in themselves valuable for education, but rather lend their value from enabling the achievement of educational aims. Secondly, that how AI can be of value in a specific educational context is dependent on what the people involved in that context value. In Sections 2, 3 we will ground our way of thinking through a conceptual analysis of the literature. To illustrate how we translate this way of thinking and its underlying premises to the context of educational assessment at Cito, we then present three case studies of AI for educational measurement in Section 4. We conclude with a discussion of our findings and a reflection on the limitations of this work.

2 Background: education and educational measurement

What we are trying to achieve with education, and what constitutes good education, can be understood through the lens of three overarching aims: qualification, socialization, and subjectification (Biesta, 2015). Qualification relates to the acquisition of knowledge and skills for reasoned and effective action. Socialization concerns the adoption of attitudes aligned with cultural norms and practices. Subjectification regards the cultivation of agency to critique and transcend cultural norms and practices. These aims of education are mutually constitutive. Qualification without socialization produces hollow competence, socialization without subjectification reduces learning to compliant replication, and subjectification without qualification yields naive agency.

Yet, educational aims are neither fixed nor universal. Educational goals reflect the values of a culture and its historical context, as well as the specifics of an institution's vision, a subject's characteristics, and a teacher's beliefs. It is up to the teacher to translate these values and specifics into concrete interactions with their students; these interactions should be designed in a way that educational goals, activities, and assessment are aligned (Biggs, 1996). Misalignment fractures education into disjointed acts where tasks and assessments lack an underlying purpose (Siegel, 2004).

For such educational aims to take shape in practice, the teacher-student relationship must function as a reliable foundation for learning and assessment. The interaction between teacher and student is driven by their shared social contract; a mutual commitment with shared expectations, responsibilities, and trust (UNESCO, 2021). For educational measurement, this social contract pertains to the student promising to deliver honest work and the teacher promising to deliver an honest assessment of that work. Recall that educational measurement is the process of inquiry by which we gain insight into a student's KSAs, which are hidden from direct observation. This inquiry is not a passive observation but a dynamic process of making and justifying claims (Wyatt-Smith and Gunn, 2009). Through this process, measurement can serve different educational purposes: certify qualification through assessment of learning, refine socialization through assessment for learning, and instigate subjectification through assessment as learning (Earl, 2003). When either side of the social contract is breached, this undermines the process of inquiry, meaning accurate assessment of KSAs and meaningful achievement of educational goals become impossible. This serves to illustrate that educational measurement is a moral act as much as a technical or practical one.

Central to the act of educational measurement are assessment tasks, which are transactional encounters designed to elicit evidence about KSAs (Dewey and Bentley, 1949). Such tasks are never neutral. Firstly, they embed assumptions about which competencies matter and how they should be demonstrated. Secondly, the interpretations of the evidence, the residue of a student's engagement with these tasks, is also not a neutral act. Linking this evidence to KSAs requires making inferences, which should be warranted for them to be of any value (Arieli-Attali et al., 2019). Statistical models formalize these inferences, but even these are not neutral, since they encode assumptions about how KSAs develop and interact. When the theoretical lens of a model misrepresents the inference it is formalizing, we risk undermining the validity of the inference.

Validity can be thought of as the degree to which what is measured corresponds to what was intended to be measured. Validity is a multi-faceted concept that can be conceptualized via a narrow or unified view. When conceptualized narrowly, validity encompasses "types of validity" such as construct validity, criterion validity, and content validity. When treated as a unitary concept, validity encompasses all facets that are relevant to answering the question whether we are measuring what we intended to measure (Messick, 1989; Kane, 2013). Validity then includes facets such as usefulness, fairness, meaningfulness and trustworthiness; facets that are not included in the narrow conceptualization. In this paper, we treat validity as a unitary concept.

The question remains as to what the role of AI is in facilitating good education through enhanced validity of educational measurement. In the next section, we address this question by tracing the way AI systems impact the structures that govern how we assign meaning to assessment evidence.

3 Method: tracing the value of AI for education

For a worthwhile discussion of the value of AI for education, it helps to work from a common definition of what AI is. Perhaps the definition of AI that will most influence education in the coming years, at least within the European Union (EU), is the definition of

AI systems outlined in the EU AI Act. The EU AI Act defines an AI system as:

> "A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." [European Union, 2024, p. 46, Article 3 (1)]

In addition to providing us with a perspective on what an AI system is, this definition also highlights the characteristics of an AI system that can be altered to achieve different results in educational contexts. We can understand the characteristics of an AI system as dials that we can tune to alter our educational assessments. Choices regarding the student and teacher data that the system requires as input (Williamson et al., 2020) and the level of autonomy of an AI system (Molenaar, 2022) are examples of such dials. They represent choices that determine how an AI system will alter the structures that govern the way we assign meaning to an assessment's resulting evidence. To determine the value of AI systems for education, we must have a complete picture of their impact on these underlying structures. This highlights the necessity of being able to specify the configuration of an AI system's dials.

In Section 4, we evaluate the value of several AI systems for education, using three core elements of educational measurement as a basis for our argument. The first of these elements is the educational measurement context. A prerequisite for developing meaningful assessments is to be able to specify your educational aims (Biesta, 2015), the intended use of the assessment (Earl, 2003), and the KSAs you wish to measure (Arieli-Attali et al., 2019). Specifying the purpose of assessments may even lead to the realization that there is not a single purpose, but rather a plurality of purposes (Newton, 2017).

The second element is the validity conceptualization used in the educational measurement context. You must be able to answer questions relating to how you intend to measure KSAs, how the AI system will influence the measurement process, and how you prioritize different validity criteria within your conceptualization of validity (Messick, 1989). By prioritization we do not mean that particular validity facets should always take precedence over others, as this is generally seen as a problematic way of approaching assessment (van Haastrecht et al., 2024). Rather, we mean that it is important to know which facets you want to put more or less emphasis on Williamson et al. (2012).

The third element is the social contract between student and teacher (UNESCO, 2021). It is essential to reflect on who is producing what work and what the impact of the introduction of the AI system is on the relationship between student and teacher. If an AI system meddles with a student's promise to deliver honest work or a teacher's promise to deliver an honest assessment of that work, this raises fundamental questions regarding what we are trying to achieve with our educational assessments. If we are to determine whether a specific AI intervention contributes to the validity of educational measurement, we must be able to answer questions such as how the AI system impacts the alignment between students and teachers with respect to educational purpose, and how differences in understanding of the AI system influence their perception of the assessment (Williamson et al., 2012). We believe that these questions are insufficiently emphasized within traditional frames of thinking about validity, such as Messick's unitary conceptualization and Kane's argument-based approach. Hence, we choose to include the social contract as a separate core element.

Having a precise description of how these elements are affected by the introduction of an AI system is a prerequisite for determining the value contribution of AI. This value contribution then corresponds directly to the degree to which the AI system supports assessment validity, in turn enabling the achievement of educational goals. Realize that this places restrictions on which AI systems can be valuable. For example, AI systems that are not interpretable or explainable will impact the shared understanding between students and teachers in ways we cannot foresee. When such systems cause misalignment between students and teachers regarding educational aims, they will generally not contribute to valuable education. Using AI beyond the limits suggested in this section runs the risk not only of obscuring student KSAs, but also of fracturing the social contract between students and teachers that is so central to meaningful education. Using AI within these limits will open up possibilities to enhance validity and shape more valuable education.

4 Case studies

To illustrate the way of thinking based on the three elements specified in the previous section, we will discuss three AI applications that were developed at our institution. For each case, we will describe the measurement context, our validity conceptualization, and our view on the social contract between student and teacher. We will explain why we believe our AI system design would aid valid measurement. For each case, we will reflect on how the way of thinking introduced in this paper is helpful in implementing AI in educational measurement. Moreover, we will illustrate how people in different educational contexts may have different views on whether these AI systems truly contribute value.

4.1 Item construction

The process of constructing assessment items is traditionally labor-intensive and time-consuming. For instance, constructing a single test with around 35 questions for the Dutch central exams at the end of secondary school generally takes more than a year and requires a team of multiple subject and testing experts. In this context, the utility of AI becomes apparent. Generative AI models, such as large language models, possess the ability to rapidly produce questions across diverse topics, simplifying the initial phase of item creation.

To try to overcome this labor intensive and time-consuming process, we collaborated in an international project to explore whether and how generative AI can help in the process of generating items for central exams, with an open AI assistant (Kruis et al., 2024). This exploration has highlighted significant deficits in the quality of AI-generated items. Initial outputs frequently failed

to meet quality benchmarks, necessitating iterative refinements and adjustments to achieve acceptable standards. This research reveals a critical issue: while AI can swiftly produce items with superficial validity, the depth and reliability required for effective assessment are often lacking. To mitigate these shortcomings, solutions for AI item generation could focus on combining generative AI with rigorous quality assurance protocols. By implementing stringent checks on AI-generated material, such tools have the potential to enhance the quality of assessment items and to shift focus toward complex matters within the construction process.

In terms of *measurement context*, clearly defining what we want to achieve with our assessment is essential. The Dutch central exams are an example of a high-stakes summative assessment where the explicit goal is assessment of learning. This introduces vastly different constraints on the integration of AI systems than would be the case for low-stakes assessments, and explains why for now the most likely way forward is a situation where humans always remain in control. In the end, efficiency is at best a secondary purpose, and we would never sacrifice trust in the quality of our exams for more efficient item construction.

The *validity conceptualization* employed for the Dutch central exams has a relatively high emphasis on reliability, fairness, and trustworthiness. Although there are some doubts regarding the quality of AI-generated items and the inclusivity of its language (Kruis et al., 2024), these seem to be hurdles that are possible to overcome with new algorithmic developments and human oversight. However, we should realize that this does not immediately imply that students, teachers, and society will trust that quality standards have not dropped. With trust being so easy to lose and so hard to gain, the facet of trustworthiness seems to provide the critical challenge for AI interventions aiming to improve the validity of centralized exams.

The *social contract* between student and teacher plays a relatively minor role in centralized, summative assessment. However, a large part of the trust in the quality of the Dutch central exams that has been built up over the years is the result of involving teachers in the process of item construction. Teachers are trusted by students and parents, and are much better than current AI systems at formulating questions that align with students' worldview. We should realize that by introducing an AI system into the equation we inevitably change the role of teachers, and thus change the foundation on which trust was built up. This is not to say that AI systems can never have a role within this measurement context, but we should be mindful of how easy it is to gamble away trust by not being thoughtful about our AI interventions.

This case study highlights an interesting dilemma concerning the implementation of AI systems for item construction. Centralized exams may offer the educational measurement context with the greatest potential for efficiency gains, but it is often considered to be ill-advised to automate high-stakes assessments with AI. However, there are many centralized or standardized assessments that are not high-stakes, often taking the form of formative assessments at earlier ages. The measurement context, validity conceptualization, and social contract for such assessments may be much better positioned for valuable automation of item construction with AI. Furthermore, in time developments in the field of generative AI could improve quality of AI-generated items and reduce validity issues.

4.2 Assessing written work

Generative AI tools have made it trivial for students to produce written work without requiring a certain level of writing ability or understanding of the subject matter. Given the pressure to perform that many students experience, the temptation to secure quick wins through AI-generated work is understandable.

In response, a proliferation of detection tools has emerged, designed to identify AI-generated content in student submissions. Yet, these tools have significant limitations: they often yield false positives, by mistakenly flagging legitimate student work as AI-generated. This can, for example, occur when students use common phrases. Additionally, students familiar with AI can iteratively modify the AI output to avoid detection. While such workarounds may provide short-term gains, they ultimately hinder students from developing writing skills and achieving meaningful learning. Furthermore, the argument that AI makes such skills obsolete is contradicted by evidence demonstrating that learning to write confers a range of broader cognitive and attitudinal benefits that extend beyond the mastery of writing alone (Curtis et al., 2019; Nuckles et al., 2020).

This background led to the creation of the SchrijfBlik prototype (Poell et al., 2025). SchrijfBlik allows teachers to create and assign writing tasks, which students can complete with the help of an integrated AI chatbot. The chatbot not only generates content upon request, but also encourages students to think critically about their assignments. The application logs key student actions, such as questions asked to the AI and copy-pasting behavior. Teachers can view a timeline that details the student's writing process and how their work developed over time. This log helps teachers identify which parts of the text were produced by the student, as illustrated in Figure 1, allowing for more targeted feedback and assessment.

In terms of the *measurement context*, our findings indicate that teachers mainly value SchrijfBlik for formative assessments. They believe that the process data enables them to give more targeted feedback on students' writing abilities. One reason teachers feel that formative assessments are more suitable for SchrijfBlik is that summative assessments typically occur in controlled classroom environments where the use of AI is restricted. In cases where a teacher's sole focus is on classroom assessment of writing ability, SchrijfBlik adds little value. However, if a teacher would be interested in assessing other KSAs, such as AI literacy or the ability to write collaboratively with AI, SchrijfBlik is better equipped to make a meaningful contribution.

Regarding *validity conceptualization*, the prioritization of different facets of validity, such as construct validity and authenticity, will rely heavily on the measurement context. When the goal is to evaluate a student's independent writing skills, SchrijfBlik strengthens construct validity by making it possible to distinguish between student-generated and AI-generated content. For constructs like collaborative writing with AI, SchrijfBlik not only improves construct validity but also provides a more authentic assessment environment, as the AI chatbot is naturally integrated into the student's workflow. On the other hand, we have found that integrating a chatbot in the writing environment encourages students to use AI more than they might normally do. Especially when teachers are looking for an isolated assessment of student writing ability, SchrijfBlik may not be the right environment. This

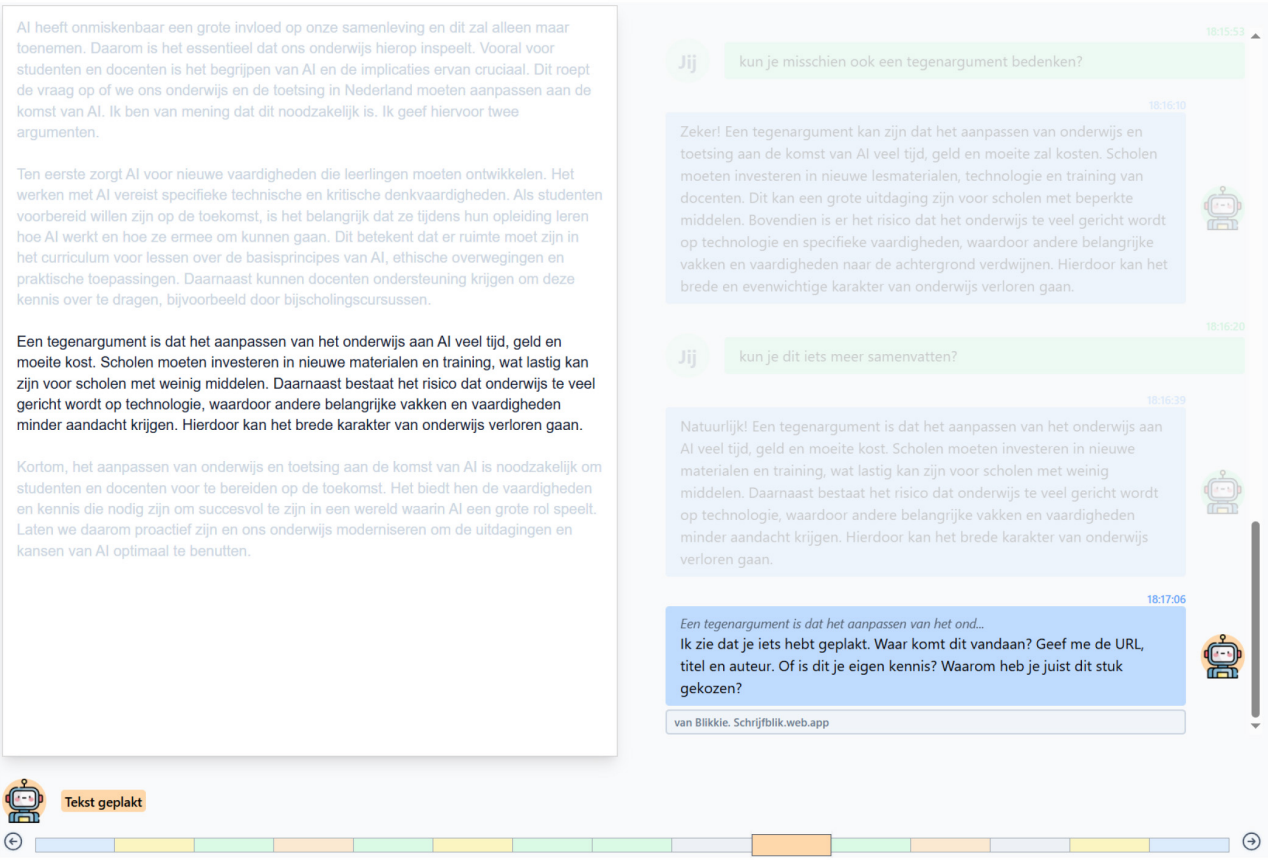

FIGURE 1
Teacher view of the SchrijfBlik prototype. The system detected a copy-paste action by the student, triggering the chatbot to ask the student to explain where they retrieved the text.

setup could not only obscure evidence of independent writing, but also create unwanted disparities between students with varying levels of AI literacy.

How we view the effect of Schrijfblik on the *social contract* between teachers and students will depend on our underlying values. On the one hand, SchrijfBlik removes any ambiguities about which part of the produced work can be attributed to the student and provides teachers with a transparent view of the student's writing process, enabling a conversation between teacher and student that is better grounded in reality. On the other hand, the use of SchrijfBlik can be seen by students as an admission by the teacher that they do not trust the student to deliver honest work. The teacher apparently feels they need to infringe on student privacy and autonomy to restore a valid process of inquiry. The exact balance between transparency, trust, privacy, and autonomy will depend on the aims and values of a particular measurement context.

In conclusion, the use of AI tools in writing assignments makes it essential to clarify the roles of both students and AI in the creation process. Tools like SchrijfBlik can offer valuable support for assessment by providing teachers with detailed insights into the writing process and the respective contributions of students and AI, especially in formative assessment. Its ability to distinguish between student and AI-generated text enhances construct validity, especially when assessing collaborative skills and AI literacy. If education professionals want to integrate AI in writing assessments, they need to discuss with their students why they have chosen this application and carefully explain the process in order not to undermine the social contract.

4.3 Grading assistance

Grading student work is, like any human activity, error-prone. In the context of the Dutch central exams, it is not uncommon to see variability between raters, even for identical student responses. Thus, we can ask whether AI could provide value by supporting a more consistent grading process. Not by replacing teachers in the grading process, but by supporting them to achieve greater consistency and fairness in their judgments. An illustrative example is the CheckMate prototype shown in Figure 2, developed in collaboration with Nationaal OnderwijsLab AI (NOLAI), a publicly funded Dutch initiative that brings together educators, researchers, and industry to co-design and validate AI-based educational innovations for primary and secondary schools. CheckMate employs natural language processing to analyze open-ended student responses, clustering those that are semantically

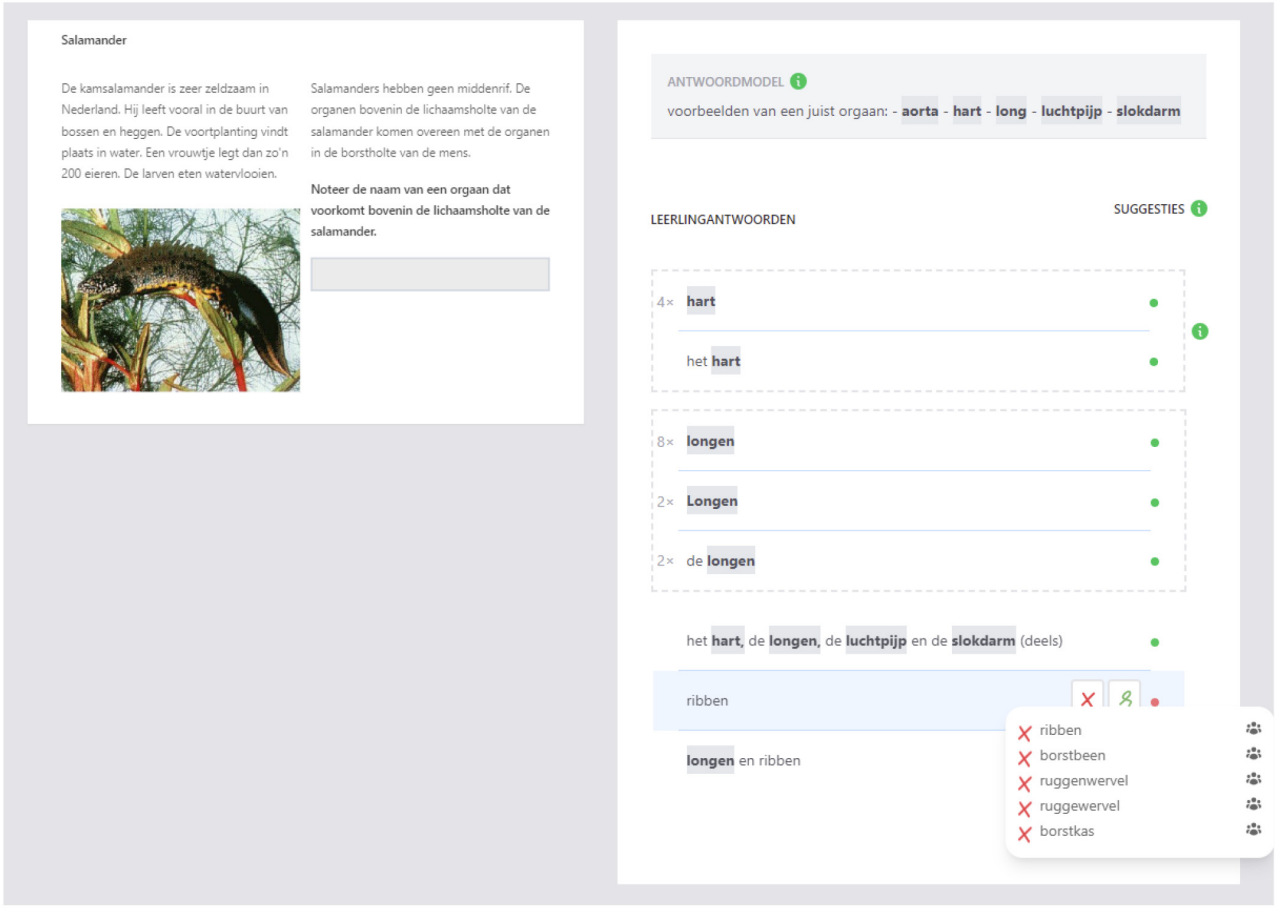

FIGURE 2
A view of the CheckMate digital grading environment with natural language processing features.

similar or even identical. These clusters are then presented to teachers, along with information on the scores assigned by other teachers to similar responses. By grouping responses, CheckMate makes it easier for teachers to apply consistent scoring criteria, to calibrate their judgments with those of their peers, and to reduce inadvertent inconsistencies.

The *measurement context* is a dominant factor in determining whether CheckMate can be of value. When grading hundreds or thousands of students on short, open-answer items in a summative assessment, clustering responses and labeling terms from an assessment template can help teachers save time and grade more consistently. However, when a teacher seeks to conduct a formative assessment with items that require longer open-ended answers or essays, CheckMate's current functionalities are practically useless. Such assessments typically require more nuanced and personalized feedback that goes beyond simply identifying surface-level similarities between responses. Teachers must consider aspects such as depth of reasoning and the way an argument is constructed, which cannot be captured by clustering algorithms. In fact, when the goal is to provide personalized feedback to students, grouping students' answers could even do more harm than good, as it may obscure important individual differences and prevent teachers from addressing specific learning needs. Whether an assessment is high-stakes or low-stakes is also an important consideration, as it determines how appropriate different levels of autonomy for the AI system are (Molenaar, 2022).

Concerning the *validity conceptualization*, CheckMate is designed to be most valuable in contexts where inter-rater reliability and consistency are prioritized. Whether CheckMate leads to more fair assessments is debatable. It could support fairness by limiting the personal biases of teachers. However, it could also introduce biases by employing a narrow definition of what constitutes the right answer, which could disadvantage e.g. minorities or non-native speakers. If this narrow vision of an AI model is enforced too strictly, the model could label even the slightest deviations from an assessment template as incorrect. This could then create a washback effect provoking students to provide answers that are to the AI system's liking, rather than providing the right answers (Filighera et al., 2024). Within the CheckMate context, this risk is counteracted by always working with a teacher-in-the-loop, but this does require the teacher to employ a critical attitude toward AI suggestions.

As AI becomes integrated into the grading process, the *social contract* between students and teachers is also affected. Although grading is often a time-consuming and repetitive task, it is a task that helps teachers to better understand their students. One

can wonder whether the efficiency and consistency gained by AI-assisted grading are worth sacrificing this understanding. Students can also be affected by the insertion of AI into the grading process, and this extends beyond the washback effect mentioned previously. Students put in effort to complete an assessment task and generally expect teachers to contribute their fair share of work to assess the student's product. When teachers start cutting corners by delegating tasks to AI systems, this can be detrimental to student motivation. Collaborating with students and teachers in the process of designing and implementing tools such as CheckMate is essential if we want to avoid this and uphold the social contract.

In summary, while AI-powered tools like CheckMate can enhance grading consistency and efficiency, their value is highly dependent on the measurement context in which they are applied. In high-stakes or complex assessments that require nuanced judgment, reliance on AI may compromise fairness or validity. In contrast, for routine or objective grading tasks, such tools can offer significant benefits. It is crucial that educators remain actively involved in selecting when and how AI is used for grading, ensuring that these tools support sound educational practices and relationships.

5 Discussion

In this paper, we introduced a way of thinking about educational measurement to clarify where AI systems can be a helpful aid, and where they may cause more harm than good. This way of thinking centers around three core elements of educational measurement that may be impacted by AI systems: the educational measurement context, the conceptualization of validity, and the social contract between student and teacher. Where traditional frames of thinking about validity, such as Messick's unitary conceptualization and Kane's argument-based approach, tend to focus on the first two core elements, AI systems also impact the relationship between students and teachers. By explicitly incorporating the social contract in our way of thinking, we aimed to highlight the necessity of considering this element when designing valuable assessments. We positioned our way of thinking as value-neutral, allowing those who use it to apply their own values based on their educational context. To demonstrate its practical usefulness, we examined three case studies. These case studies illustrate how our way of thinking can help to clarify the benefits and challenges of integrating AI in educational measurement.

In the case studies, we discussed how efficiency gains, which are often presented as a benefit of using AI systems, are at best a secondary purpose in the context of education. In the item construction case, the conceptualization of validity was the most important element to consider. We concluded that trust in exam quality is key. This trust may not be sacrificed for more time efficient construction. In the second case study covering writing assignments, the social contract was the central element. We laid bare the frailty of the social contract between teacher and student when the teacher admits they do not trust the student to deliver honest work. Finally, in the AI-assisted grading case study, we showed how clarity regarding the measurement context (specifically the type of assessment) is vital in determining whether tools like CheckMate can add value. These insights, which followed through the application of our way of thinking, need careful consideration. Our way of thinking enables us to critically examine both the advantages and limitations of AI tools, creating space for a balanced discussion.

While our approach provides a guided way of thinking about the impact of AI systems on educational measurement, it also has limitations. Firstly, by consciously choosing to stay away from a more concrete framework, the way of thinking presented in this paper is not a simple checklist. Instead, it is intentionally designed to be adaptable rather than prescriptive, allowing for broad applicability and flexibility, though this may make it less straightforward to apply in practice. Because we do not provide specific guidelines on what educators should do or which regulatory requirements and data privacy concerns they should consider, this flexibility allows the approach to be tailored to the diverse needs and values of different educational contexts. We hope that the discussion of case studies in Section 4 has provided an indication of how this way of thinking can be employed effectively. Secondly, we realize that by treating validity as a unitary concept and by avoiding the introduction of normative elements, our way of thinking does not align with views that lean toward specifying right and wrong ways to conduct educational measurement.

The three case studies in Section 4 show that, based on our own beliefs and values, we do not believe all approaches to educational measurement are equally valuable in different situations. However, in our way of thinking, we wish to recognize that there are always choices to be made in which validity facets to prioritize. We believe that these choices should not be normatively prescribed by external parties but should be made collaboratively with all stakeholders involved in an educational context. In this age of artificial intelligence, it is perhaps more necessary than ever to continually and collaboratively reflect on the purpose of education with relevant stakeholders. The way of thinking along the three elements introduced in this paper can facilitate critical reflection on what we value in education. It also shapes how we wish to design our educational assessments and how we can effectively incorporate AI in the assessment process.

Author contributions

MH: Conceptualization, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. LG: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. MJ-P: Methodology, Writing – review & editing. FB: Methodology, Writing – review & editing. JK: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Funding

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

Acknowledgments

We want to thank our colleagues at CitoLab for their contributions to the prototypes discussed in this work, and for their constant flow of ideas to inspire this article.

Conflict of interest

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

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI was used in the ideation phase of the writing process (ChatGPT-4o and DeepSeek-R1) and was used to provide suggestions for improvements to existing text (Writefull suggestions in Overleaf). The author(s) take full responsibility for the use of generative AI in the preparation of this manuscript.

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

Publisher's note

References

Aloisi, C. (2023). The future of standardised assessment: validity and trust in algorithms for assessment and scoring. *Eur. J. Educ.* 58, 98–110. doi: 10.1111/ejed.12542

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., and von Davier, A. A. (2019). The Expanded Evidence-Centered Design (e-ECD) for learning and assessment systems: a framework for incorporating learning goals and processes within assessment design. *Front. Psychol.* 10. doi: 10.3389/fpsyg.2019.00853

Biesta, G. J. J. (2015). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. New York, NY: Routledge.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Educ.* 32, 347–364. doi: 10.1007/BF00138871

Curtis, P. R., Kaiser, A. P., Estabrook, R., and Roberts, M. Y. (2019). The longitudinal effects of early language intervention on children's problem behaviors. *Child Dev.* 90, 576–592. doi: 10.1111/cdev.12942

Delandshere, G. (2002). Assessment as inquiry. *Teach. College Record*. 104, 1461–1484. doi: 10.1111/1467-9620.00210

Dewey, J. (1910). *How We Think*. Lexington, MA: D.C. Heath & Company.

Dewey, J., and Bentley, A. F. (1949). *Knowing and the Known*. Boston, MA: Beacon Press.

Earl, L. M. (2003). *Assessment As Learning: Using Classroom Assessment to Maximize Student Learning*. Los Angeles, CA: SAGE Publications.

European Union (2024). "AI act: regulation 2024/1689 of the European Parliament and of the council," in *Regulation, Official Journal of the European Union* (Brussels: European Union).

Filighera, A., Ochs, S., Steuer, T., and Tregel, T. (2024). Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. *Int. J. Artif. Intellig. Educ.* 34, 616–646. doi: 10.1007/s40593-023-00361-2

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educ. Measurem.* 50, 1–73. doi: 10.1111/jedm.12000

Kruis, J., Pera, M. S., Napel, Z., t., Landoni, M., Murgia, E., et al. (2024). "Toward personalised learning experiences: beyond prompt engineering," in *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (New York, NY: Association for Computing Machinery), 644–649.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Research.* 18, 5–11. doi: 10.3102/0013189X018002005

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *Eur. J. Educ.* 57, 632–645. doi: 10.1111/ejed.12527

Newton, P. E. (2017). There is more to educational measurement than measuring: the importance of embracing purpose pluralism. *Educ. Measurem.* 36, 5–15. doi: 10.1111/emip.12146

Nuckles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., and Renkl, A. (2020). The self-regulation-view in writing-to-learn: using journal writing to optimize cognitive load in self-regulated learning. *Educ. Psychol. Rev.* 32, 1089–1126. doi: 10.1007/s10648-020-09541-1

Poell, T., Maas, L., Balk, M., and van Haastrecht, M. (2025). "SchrijfBlik: safeguarding the validity of writing assessment in the age of AI," in *Proceedings of the 26th International Conference on Artificial Intelligence in Education*, eds. A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, and S. Isotani (Cham: Springer Nature Switzerland), 71–80.

Siegel, H. (2004). High stakes testing, educational aims and ideals, and responsible assessment. *Theory Res. Educ.* 2, 219–233. doi: 10.1177/1477878504046515

UNESCO (2021). "Reimagining our futures together: a new social contract for education," in Technical *report, Educational and Cultural Organization of the United Nations* (Paris: UNESCO).

van Haastrecht, M., Haas, M., Brinkhuis, M., and Spruit, M. (2024). Understanding validity criteria in technology-enhanced learning: a systematic literature review. *Comp. Educ.* 220:105128. doi: 10.1016/j.compedu.2024.105128

Williamson, B., Bayne, S., and Shay, S. (2020). The datafication of teaching in Higher Education: critical issues and perspectives. *Teach. Higher Educ.* 25, 351–365. doi: 10.1080/13562517.2020.1748811

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educ. Measurem.* 31, 2–13. doi: 10.1111/j.1745-3992.2011.00223.x

Wyatt-Smith, C., and Gunn, S. (2009). "Towards theorising assessment as critical inquiry," in *Educational Assessment in the 21st Century: Connecting Theory and Practice*, eds. C. Wyatt-Smith, and J. J. Cumming (Dordrecht: Springer Netherlands), 83–102.