frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 07 October 2024
DOI 10.3389/frai.2024.1460651

OPEN ACCESS

EDITED BY
Rita Orji,
Dalhousie University, Canada

REVIEWED BY
Leonardo Brandão Marques,
Federal University of Alagoas, Brazil
Grace Ataguba,
Academic City University College, Ghana

*CORRESPONDENCE
Burcu Arslan
barslan@ets.org

RECEIVED 06 July 2024
ACCEPTED 04 September 2024
PUBLISHED 07 October 2024

CITATION
Arslan B, Lehman B, Tenison C, Sparks JR, López AA, Gu L and Zapata-Rivera D (2024) Opportunities and challenges of using generative AI to personalize educational assessment.
Front. Artif. Intell. 7:1460651.
doi: 10.3389/frai.2024.1460651

Opportunities and challenges of using generative AI to personalize educational assessment

Burcu Arslan*, Blair Lehman, Caitlin Tenison, Jesse R. Sparks, Alexis A. López, Lin Gu and Diego Zapata-Rivera

ETS Research Institute, Princeton, NJ, United States

In line with the positive effects of personalized learning, personalized assessments are expected to maximize learner motivation and engagement, allowing learners to show what they truly know and can do. Considering the advances in Generative Artificial Intelligence (GenAI), in this perspective article, we elaborate on the opportunities of integrating GenAI into personalized educational assessments to maximize learner engagement, performance, and access. We also draw attention to the challenges of integrating GenAI into personalized educational assessments regarding its potential risks to the assessment's core values of validity, reliability, and fairness. Finally, we discuss possible solutions and future directions.

KEYWORDS

personalization, educational assessment, generative artificial intelligence, validity, reliability, fairness

1 Introduction

Personalized learning has been shown to enhance learner motivation, engagement, and performance (Bernacki et al., 2021; Walkington, 2013; Walkington and Bernacki, 2018, 2019). Personalization can be delivered via humans (e.g., students or teachers), digital assessment systems (e.g., via a virtual agent embedded in a digital platform), or a combination (e.g., recommender systems). In educational assessment, standardization has been one of the most essential requirements for fair and valid measurement (Sireci, 2020). However, more recent discussions put the learners in front and expect that personalized assessments yield similar benefits to personalized learning (Bennett, 2023; Buzick et al., 2023; Sireci, 2020). The transition from standardized to more personalized assessment of learning (i.e., summative assessment) and assessment *for* learning (i.e., formative assessment) comes with inherent challenges in ensuring the validity, reliability, and fairness of more tailored, individualized assessments.

Artificial intelligence (AI) in education dates back more than four decades (see Holmes and Tuomi, 2022; Williamson and Eynon, 2020, for reviews). However, recent technological advancements and generative AI (GenAI) have broadened AI's scale and potential applications in education due to its ability to create human-like text, being generic enough to be employed for different tasks, and real-time personalization capabilities. GenAI is a subcategory of AI designed to generate content, including images, videos, and text. Large language models (LLMs) are specifically trained on vast amounts of text data. When powered by LLMs, GenAI models can have a contextual understanding, enhanced memory, and create content based on natural language input (Hadi et al., 2023).

Recent research has increasingly focused on the integration of GenAI and LLMs in educational settings, examining their potential (e.g., Barany et al., 2024; Gökoğlu, 2024; Hu, 2023; Kasneci et al., 2023; Mazzullo et al., 2023; Nguyen et al., 2023; Olney, 2023; Pankiewicz and Baker, 2023; Pardos and Bhandari, 2024; Wang et al., 2022). Similarly, some studies focus on the application of GenAI and LLMs in educational assessment, exploring their impact and implications (e.g., Bulut et al., 2024; Hao et al., 2024; Jiang et al., 2024; von Davier, 2023; Swiecki et al., 2022).

Despite these advancements, to our knowledge, the potential opportunities and challenges of using GenAI to personalize educational assessment have not been explored. As we mentioned above, the shift from one-size-fits-all assessments to more culturally relevant and responsive approaches is becoming more critical, especially as stakeholders recognize the limitations of traditional assessments in responding to the needs of diverse populations. Thus, personalized educational assessments are increasingly viewed as a means to enhance learner engagement, performance, and access (Bennett, 2023; Buzick et al., 2023; Randall et al., 2022; Sireci, 2020).

Similar to the other application areas, advances in GenAI offer opportunities and challenges to personalized educational assessment (see Kirk et al., 2024, for benefits and risks of personalization in general with LLMs). GenAI can be integrated with the existing frameworks for including personalization, adaptation, or responsiveness in assessments, such as caring assessments (Lehman et al., 2024; Zapata-Rivera et al., 2020), socioculturally responsive assessments (Bennett, 2023), formative assessments (Bennett, 2011; Black and Wiliam, 2009), and intelligent tutoring systems (Corbett et al., 1997; Graesser et al., 2012). For example, in line with the caring assessment framework, GenAI may be leveraged to tailor content to the learner's emotional, motivational, and cognitive state. Similarly, in line with socioculturally responsive assessments, GenAI may adapt assessment content to reflect diverse perspectives and contexts, considering the learner's cultural background. Moreover, GenAI may enhance formative assessments and intelligent tutors by providing real-time, personalized feedback in a conversational style that helps learners improve continuously (e.g., Cheng et al., 2024). By leveraging these established frameworks, GenAI can offer robust personalized assessments that are not only effective but also responsive to the diverse needs of learners.

Integrating GenAI into the existing frameworks may play a crucial role in efficiently personalizing educational assessments by automatically generating images, videos, scenarios, and metadata and evaluating and scoring assessment items. Moreover, GenAI has the potential to generate or modify assessment items in real-time (Arslan, 2024), adapt to the learner's responses, performance, interests, or cultural background, and provide personalized feedback and reporting dashboards. Additionally, GenAI can be used to create personalized conversations about the construct that can be used for assessment purposes to create assessment content at varying levels of language complexity or translate it into multiple languages. These potential uses of GenAI can help to achieve the previous efforts of enhancing the assessment experience through maximizing learner performance and engagement, activating existing *funds of knowledge* (González et al., 2005), and making assessments more relevant and accessible to learners, including neurodiverse and multilingual learners (Sireci, 2020).

However, using GenAI to personalize educational assessment also introduces significant challenges, such as ensuring fairness and maintaining validity and reliability. Research increasingly highlights the challenges and risks associated with GenAI, including issues such as bias, copyright infringement, the potential for harmful content, minimal control over its output, security concerns, and lack of interpretability and explainability (Bender et al., 2021; Kasneci et al., 2023). Table 1 shows the potential opportunities and challenges of using GenAI for personalized assessments, potential solutions, and future directions.

2 Potential opportunities for applying GenAI in personalized assessments

GenAI may offer significant opportunities to enhance the personalization of assessments to maximize learner motivation and engagement, performance, and access.

2.1 Personalization for maximizing motivation and engagement

Increased motivation during test-taking leads to cognitive engagement, resulting in learners giving their best effort when answering assessment items (Finn, 2015; Wise and Kong, 2005; Wise, 2017). Cognitive engagement improves the likelihood that test scores will accurately represent what learners know and can do, as the interpretation of scores relies on the assumption that learners are trying their best (Finn, 2015; Wise, 2017). An effective way of maximizing engagement for learners with diverse interests and sociocultural is to personalize the context of assessment items (Bennett, 2023; Bernacki and Walkington, 2018; Sireci, 2020; Walkington and Bernacki, 2018). Context personalization can significantly enhance learner motivation and engagement by allowing learners to bring their cultural identity to the learning environment, leading to better learning outcomes (Walkington, 2013; Walkington and Bernacki, 2018, 2019).

LLMs have made it possible to personalize the context of assessment items *during* assessment based on each learner's input about their interests embedded in their cultural identities, thus maximizing engagement through situational interest (see Hidi and Renninger, 2006) and has the potential to allow learners to show what they know and can do (Arslan, 2024). Unlike personalization approaches that leverage background variables (e.g., race/ethnicity) to create culturally relevant forms and assign each form to a group of learners based on their demographic information (e.g., Sinharay and Johnson, 2024), using LLMs offers real-time tailoring of content to individual interests and cultural background by providing learners agency and relevance that is often missing in standardized assessments. This approach acknowledges the diversity among learners, avoiding the pitfall of assuming homogeneity within groups (Arslan, 2024).

Integrating conversational virtual agents into assessment platforms is another way of making assessments more engaging and user-friendly. The virtual agents, powered by LLMs, can respond to queries in natural language, providing real-time, contextual support that assists both students and teachers during their interactions with

TABLE 1 Potential opportunities and challenges of using GenAI for personalized assessments, potential solutions, and future directions.

Opportunities	Challenges	Consequences	Potential solutions and future directions
Maximizing Engagement • On-the-fly context personalization (e.g., Arslan, 2024; Bennett, 2023; Bernacki and Walkington, 2018; Hidi and Renninger, 2006; Sireci, 2020; Walkington, 2013; Walkington and Bernacki, 2018, 2019). • Conversational agents to enhance usability (Zapata-Rivera, 2012; Bull and Kay, 2016; Zapata-Rivera and Greer, 2002). Maximizing Performance • Conversational agents to gather additional evidence, to provide just-in-time feedback (e.g., Kochmar et al., 2020; Ma et al., 2014; Mazzullo et al., 2023; Meyer et al., 2024; Matelsky et al., 2023; Pardos and Bhandari, 2024; Wang and Han, 2021), and to enhance dashboards for reporting (e.g., Forsyth et al., 2017; Xhakaj et al., 2017). Increasing Access • Conversational agents, scaffolds, and language supports for neurodiverse and multilingual learners (e.g., Ali et al., 2020; López, 2023; Yang, 2024).	• Lack of control over the quality and content of the output (e.g., Bender et al., 2021; Jurenka et al., 2024). • Hallucinations, potential bias and fairness issues (e.g., Jurenka et al., 2024; Hu, 2023; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024; Ye et al., 2023; Jurenka et al., 2024). • Lack of interpretability and explainability of the system's underlying decision-making (see Jurenka et al., 2024 for a survey).	Jeopardizing assessment's core values of: • Validity. • Reliability. • Fairness (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US), 2014; Jurenka et al., 2024).	• Developing guidelines and standards for the ethical use of GenAI in personalized assessment (e.g., Hu, 2023; Jurenka et al., 2024). • Aligning the purpose and goals of the assessment with how GenAI is being leveraged (Jurenka et al., 2024). • Developing methodologies to evaluate the quality of the output of GenAI (e.g., human-in-the loop approaches; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024). • Working with practitioners and students to co-design solutions (Penuel, 2019). • Identifying and implementing guardrails (Rai et al., 2024). • Combining neuro-symbolic approaches and/or using computational cognitive architectures to create decision-making systems that leverage knowledge of human cognition (e.g., Jurenka et al., 2024; Jurenka et al., 2024).

10.3389/frai.2024.1460651

the platform (Zapata-Rivera, 2012; Bull and Kay, 2016; Zapata-Rivera and Greer, 2002).

2.2 Personalization for maximizing performance

Unlike traditional summative assessments, personalized formative assessments can significantly enhance performance by providing feedback tailored to each learner's needs (e.g., Kochmar et al., 2020; Ma et al., 2014; Mazzullo et al., 2023; Hu, 2023; Wang and Han, 2021). LLMs can generate hints and adaptive feedback during assessments at scale in an efficient way, helping learners understand their mistakes and learn from them in real-time (e.g., Meyer et al., 2024; Matelsky et al., 2023; Pardos and Bhandari, 2024) and facilitate adaptive conversations that guide learners through their thought processes (Hu, 2023; Forsyth et al., 2024; Zapata-Rivera et al., 2024).

LLMs can also enhance reporting by providing detailed, narrative insights for learners, teachers, and other interest holders. These insights can help interest holders understand assessment information more deeply and make informed decisions. For example, it can influence what teachers know about their student and their decision-making through dashboards (e.g., Forsyth et al., 2017; Xhakaj et al., 2017).

2.3 Personalization for increasing access

Personalized assessments can be crucial in increasing access for diverse learner populations, including neurodiverse and/or multilingual learners. LLMs offer various options for making assessments more linguistically responsive to the needs of multilingual learners (Yang, 2024). A significant way LLMs can enhance accessibility for multilingual learners is by providing support and scaffolds, such as translations to the learner's preferred language, language simplification, glossaries, and read-aloud features. These tools allow multilingual learners—who comprise 10.6% of the student population in US public schools (National Center for Education Statistics, 2024a)—to utilize all available linguistic resources without compromising the construct being measured. These tools offer multilingual learners alternative ways to access and engage with assessment content, ensuring that language barriers do not block learners' ability to fully demonstrate what they know and can do (Bennett, 2023; Sireci, 2020). In this context, LLMs are leveraged to provide enriched, inclusive means for all learners to access the assessment content and showcase their conceptual understanding using multiple modes of communication (e.g., linguistic, visual, aural, spatial, gestural) to reflect the diversity of needs and abilities in U.S. public schools (National Center for Education Statistics, 2024b). In essence, LLMs allow learners to use their entire linguistic repertoire, enabling them to express their KSAs through multiple forms of representation, including oral and written language and drawings (García and Wei, 2014; López, 2023). This approach, associated with translanguaging, supports providing multiple forms of expression, making assessments more inclusive and reflective of learners' diverse backgrounds (Bennett, 2023).

Conversational virtual agents powered by LLMs can also be used to further support usability for neurodiverse and/or multilingual learners by enabling interactive, natural language-based supports with a choice of spoken and written communication in understanding and navigating the assessment platform, interpreting assessment items, and providing real-time, context-sensitive assistance. (e.g., see Ali et al., 2020). This potential application makes the platform more user-friendly, as discussed in the above section, and may ensure that all learners, regardless of their language proficiency, can fully participate in the assessment process.

3 Challenges, potential solutions, and future directions

Despite its potential, GenAI introduces significant challenges for personalized assessments. In this section, we first mention GenAI's challenges in this context. Subsequently, we provide an overview of potential solutions to these challenges and future directions for research.

3.1 Challenges

Alongside research applying GenAI to new problems and domains, a growing body of work highlights the limitations and risks associated with its use. These discussions address potential biases, copywriting infringement, and the harmful content that can be introduced by large training datasets over which users have little control (Bender et al., 2021). Additionally, concerns about data privacy and security, particularly in educational contexts, are increasingly relevant when using these models (Kasneci et al., 2023). These general issues pose specific challenges when considering how GenAI can be used responsibly to support the design, administration, and reporting of personalized assessments while upholding the core values of validity, reliability, and fairness (see Johnson, 2024). Although these challenges may vary depending on the type and purpose of the assessment (e.g., formative vs. summative), we discuss several overarching challenges that are likely to shape the future development of personalized assessments using GenAI.

Personalizing assessments with GenAI offers benefits such as reducing construct-irrelevant variance indirectly by maximizing engagement (e.g., see Section 2.1) or directly by maximizing access (e.g., Section 2.3). However, without careful use, GenAI is just as likely to introduce new sources of construct-irrelevant variance. Approaches like Evidence-Centered Design (ECD; Mislevy et al., 2003) systematically align every aspect of the assessment process with theoretical and empirical evidence needed to support the claims made based on test scores. Part of the strength of this approach for generating valid assessments is the transparency at each step of the assessment development process and mapping decisions made to the intended interpretations and uses of the test. When using GenAI for on-the-fly content generation (e.g., see Section 2.1) or as a conversational virtual agent (e.g., see Sections 2.1 and 2.3), the lack of control over the output of GenAI makes it harder to ensure that the assessment content is measuring what we intend to measure (e.g., see Hong et al., 2024). With less control over the content, risks span from the introjection of inappropriate (e.g., see Greshake et al., 2023), non-sensical (Ye et al., 2023), incorrect (Hicks et al., 2024), or biased content and representations that these models have been known to

exhibit (Cheung et al., 2024; Jiang et al., 2024; UNESCO, IRCAI, 2024; Schleifer et al., 2024; Zhou et al., 2024). Moreover, LLMs perform complex computations, complicating the interpretation of their decision-making processes. This 'black-box' nature of GenAI makes it harder to detect the sources of problematic output and to create explanations for interest holders (see Zhao et al., 2024 for a survey of the explainability of LLMs).

One of the cornerstone principles of standardized summative assessments is the consistency of test forms and the comparability of scores, which ensures the reliability and validity of scores across different test administrations (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US), 2014). For example, in standardized summative assessments, on-the-fly personalization with GenAI, which generates uniquely tailored items during the assessment, may introduce construct-irrelevant variance into the measurement. This poses significant challenges to critical tenets of reliability and validity and complicates the currently established process for evaluating and documenting the reliability or precision of a given assessment. These challenges add a new dimension to ongoing discussions of the need for an expanded psychometric toolbox (such as computational psychometrics; von Davier et al., 2021), as well as more explicit guidance on valid score inferences when incorporating AI (Huggins-Manley et al., 2022) and personalization in assessments (Buzick et al., 2023).

Finally, developing and maintaining GenAI models specialized for personalized assessments involves numerous technical challenges. While prompt engineering is a popular method for adjusting a GenAI model's behavior, its ability to alter the model's actions is limited to what the model has already learned during pre-training (Bozkurt and Sharma, 2023; Jurenka et al., 2024). Alternative approaches like fine-tuning are much more expensive, requiring both quality data and expertise (e.g., Chevalier et al., 2024). Lastly, concerns regarding the hosting and management of GenAI models highlight critical data privacy and security issues. These are especially pertinent in educational contexts, where the sensitivity of learner data requires stringent security measures and ethical considerations (see Johnson, 2024).

3.2 Potential solutions and future directions

There are several key areas for future research to understand better how GenAI can be leveraged for personalized assessments. The first area is identifying, developing, and distributing guidelines and standards for the ethical use of GenAI in personalized assessments. A set of guidelines and standards helps guide future research and development and facilitates clear expectations for interest holders. Several emerging efforts exist to establish responsible AI standards in educational assessment (Burstein, 2023; Johnson, 2024). However, continued work is needed to establish guidelines and standards that encompass the full potential uses of GenAI in assessment design and development (e.g., content development to be evaluated by humans vs. on-the-fly personalization). To this end, as we briefly mentioned in the Introduction, existing frameworks for personalization, adaptation, and responsiveness in assessments may help identify these potential uses and important use cases.

The second area for future research is identifying how to best leverage GenAI in different testing contexts. As is typically the case in education, there is unlikely to be a one-size-fits-all solution for leveraging GenAI for personalized assessments (Bennett, 2023). For example, on-the-fly item generation may not be appropriate for a summative, high-stakes assessment in which there are high demands for score comparability. However, it may be appropriate for a formative, low-stakes assessment in a classroom context. Moreover, when additional approaches are taken to mitigate the inherent challenges of GenAI (e.g., nonfactual information and bias), it may be appropriate to leverage it to provide learners with conversational support during the assessment. Thus, it is essential to align the purpose and goals of the assessment with how GenAI is being leveraged and to develop methodologies to evaluate the quality of the output of GenAI before operationalizing the personalized assessments (e.g., see Zapata-Rivera et al., 2024 for leveraging ECD). It will be critical to regularly evaluate the impact of using GenAI-developed content for assessments on the perceptions of various interest holders when applied in different manners to different testing contexts. (e.g., teachers, learners, policymakers). It will also be essential to leverage GenAI to address the current needs of practitioners (e.g., teachers, assessment developers) and learners to improve the experience of developing, administering, and completing assessments. For example, teachers may struggle to provide all aspects of students' Individualized Education Programs (IEPs) during an assessment due to tools without the appropriate nuance and/or resource limitations, such as one teacher in a class of 30 students (Lehman et al., submitted).[1] Researchers can work with practitioners and students to co-design solutions to these real-world problems that utilize GenAI (Penuel, 2019).

When establishing how best to leverage GenAI in different testing contexts, a third area of research is needed to identify the guardrails that must be implemented to address some of the abovementioned challenges. While it may be tempting to let GenAI run free to maximize its potential benefits fully, key guardrails can be implemented to limit unintended negative consequences and maintain rigorous, appropriate content for personalized assessments. For example, implementing a 'human-in-the-loop' approach allows for human inspection and evaluation before GenAI-generated content is presented to learners (Amirizaniani et al., 2024; Drori and Te'eni, 2024; Park, 2024). However, this type of human review can limit some potential uses of GenAI, such as on-the-fly personalization. Moreover, rigorous research is essential to narrow the decision space for GenAI and mitigate the 'black-box' nature of LLMs. This can be achieved by integrating neuro-symbolic approaches or using computational cognitive architectures to develop decision-making systems that leverage an understanding of human cognition (e.g., Sumers et al., 2023; Sun, 2024). Additionally, combining these approaches with insights from key interest holders—such as teachers, students, and assessment developers—can help identify effective ways to utilize GenAI while minimizing unintended negative consequences.

1 Lehman, B., Gooch, R., Tenison, C., & Sparks, J. R. (submitted). The role of teachers in digital personalized assessments. Paper submitted to the annual meeting of the American Educational Research Association. Denver, CO.

The previous areas for future research have focused on the content generation process via GenAI. However, there is also a need for rigorous research to evaluate the personalized assessments that are developed with GenAI. This research will need to evaluate the quality of the content developed with or by GenAI and how the use of GenAI impacts the broader uptake of personalized assessments. When appropriate, it will be necessary to evaluate the utility of GenAI content within the current assessment development process. For example, it will be necessary to document if GenAI content results in more efficient content development processes that still maintain high levels of quality (e.g., see Park, 2024). Another area for future research is how GenAI could be leveraged to support response scoring, which could support personalized assessment and more efficient reporting (e.g., Section 2.2.).

Lastly, the full potential of utilizing GenAI for personalized assessments can only be realized if interest holders (e.g., teachers, learners, curriculum experts, and policymakers) view those assessments as valid, reliable, and fair, thus trustworthy and helpful in supporting learning.

4 Conclusion

Overall, there is a significant opportunity to enhance the deployment and effectiveness of personalized assessments, which could offer learners more relevant test materials, leading to greater engagement, improved performance, and broader access. This, in turn, has the potential to produce more valid test outcomes. However, while the potential of GenAI to create more valuable assessments is promising, it is crucial to proceed with caution. The field must continue to explore how GenAI can be effectively harnessed, but this exploration should be grounded in a rigorous evaluation of its utility.

As we move forward, it is essential not to abandon the potential for future advancements in assessments in favor of holding onto outdated development and evaluation processes (Huggins-Manley et al., 2022; Sireci, 2020). While embracing the possibilities offered by AI, we must ensure that these new tools are evaluated against criteria that recognize the affordances of both current and future technologies. However, this should never come at the expense of the core values of assessments—validity, reliability, fairness, and alignment with valued educational goals. By balancing innovation with caution, we can strive to create assessments that are both cutting-edge and trustworthy.

Author contributions

BA: Conceptualization, Writing – original draft, Writing – review & editing. BL: Conceptualization, Writing – original draft, Writing – review & editing. CT: Conceptualization, Writing – original draft, Writing – review & editing. JS: Conceptualization, Writing – original draft, Writing – review & editing. AL: Conceptualization, Writing – original draft, Writing – review & editing. LG: Conceptualization, Writing – original draft, Writing – review & editing. DZ-R: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by ETS Research Institute. This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

Conflict of interest

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

Publisher's note

References

Ali, M. R., Razavi, S. Z., Langevin, R., Al Mamun, A., and Kane, B. (2020). "A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons." In *Proceedings of the 20th ACM international conference on intelligent virtual agents* (pp. 1–8).

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

Amirizaniani, M., Yao, J., Lavergne, A., Okada, E. S., and Chadha, A. (2024). Developing a framework for auditing large language models using human-in-the-loop. arXiv preprint at: https://arxiv.org/pdf/2403.16809

Arslan, B. (2024). Personalized, adaptive, and inclusive digital assessment and learning environments. [conference presentation]. E-ADAPT conference, Potsdam, Germany. Available at: https://osf.io/82p5f/?view_only=cba3f410bc1e462fb086e3361ffed0bc (Accessed September 04, 2024).

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., and Andres, A. L. (2024). "ChatGPT for education research: exploring the potential of large language models for qualitative codebook development." In *International conference on artificial intelligence in education* (pp. 134–149). Cham: springer Nature Switzerland.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?". In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Principles, Policy & Prac.* 18, 5–25. doi: 10.1080/0969594X.2010.513678

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educ. Assess.* 28, 83–104. doi: 10.1080/10627197.2023.2202312

Bernacki, M. L., Greene, M. J., and Lobczowski, N. G. (2021). A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose (s)? *Educ. Psychol. Rev.* 33, 1675–1715. doi: 10.1007/s10648-021-09615-8

Bernacki, M. L., and Walkington, C. (2018). The role of situational interest in personalized learning. *J. Educ. Psychol.* 110, 864–881. doi: 10.1037/edu0000250

Black, P., and Wiliam, D. (2009). Developing a theory of formative assessment. *Educ. Assess. Eval. Account.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Bozkurt, A., and Sharma, R. C. (2023). Generative AI and prompt engineering: the art of whispering to let the genie out of the algorithmic world. *Asian J. Distance Educ.* 18, i–vii.

Bull, S., and Kay, J. (2016). SMILI☺: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *Int. J. Artif. Intell. Educ.* 26, 293–331. doi: 10.1007/s40593-015-0090-8

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., et al. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *arXiv preprint arXiv* 2406:18900. doi: 10.48550/arXiv.2406.18900

Burstein, J. (2023). The Duolingo English Test Responsible AI Standards. Retrieved from https://go.duolingo.com/ResponsibleAI (Accessed July 7, 2024).

Buzick, H. M., Casabianca, J. M., and Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educ. Meas. Issues Pract.* 42, 5–11. doi: 10.1111/emip.12551

Cheng, L., Croteau, E., Baral, S., Heffernan, C., and Heffernan, N. (2024). Facilitating student learning with a chatbot in an online math learning platform. *J. Educ. Comput. Res.* 62, 907–937. doi: 10.1177/07356331241226592

Cheung, V., Maier, M., and Lieder, F. (2024). Large language models amplify human biases in moral decision-making. *Psyarxiv preprint*. doi: 10.31234/osf.io/aj46b

Chevalier, A., Geng, J., Wettig, A., Chen, H., Mizera, S., Annala, T., et al. (2024). Language models as science tutors. *arXiv preprint* arXiv. doi: 10.48550/arXiv.2402.11111

Corbett, A. T., Koedinger, K. R., and Anderson, J. R. (1997). "Intelligent tutoring systems" in *Handbook of human-computer interaction, second, completely revised edition*. eds. M. Helander, T. K. Landauer and P. Prabhu (North-Holland: Elsevier Science B. V.), 849–874.

Drori, I., and Te'eni, D. (2024). Human-in-the-loop AI reviewing: feasibility, opportunities, and risks. *J. Assoc. Inf. Syst.* 25, 98–109. doi: 10.17705/1jais.00867

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Res. Report Series* 2015, 1–17. doi: 10.1002/ets2.12067

Forsyth, C. M., Peters, S., Zapata-Rivera, D., Lentini, J., Graesser, A. C., and Cai, Z. (2017). Interactive score reporting: an AutoTutor-based system for teachers. In R. Baker, E. Andre, X. Hu, T. Rodrigo and BoulayB. du (Eds.), Proceedings of the international conference on artificial intelligence in education, LNCS. Switzerland: springer Verlag. pp. 506–509.

Forsyth, C.M., Zapata-Rivera, D., Graf, A., and Jiang, Y. (2024). "Complex conversations: LLMs vs. knowledge engineered conversation-based assessment." In *Proceedings of the international conference on educational data mining*.

García, O., and Wei, L. (2014). Translanguaging: Language, Bilingualism and Education. US: Palgrave Macmillan.

Gökoğlu, S. (2024). "Challenges and limitations of generative AI in education," in *Transforming education with generative AI*. ed. N. Gunsel (GI Global), 158–181.

González, N., Moll, L. C., and Amanti, C. (2005). Funds of knowledge: Theorizing practices in households, communities, and classrooms. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Graesser, A. C., Conley, M. W., and Olney, A. (2012). "Intelligent tutoring systems" in APA Educational Psychology handbook, Vol. 3. Application to learning and teaching. eds. K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major and H. L. Swanson (American Psychological Association), 451–473.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). "Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection." In *Proceedings of the 16th ACM workshop on artificial intelligence and security* (pp. 79–90).

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., et al. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv*. doi: 10.36227/techrxiv.23589741.v2

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., and Harris, D. J. (2024). Transforming assessment: the impacts and implications of large language models and generative ai. *Educ. Meas. Issues Pract.* 43, 16–29. doi: 10.1111/emip.12602

Hicks, M. T., Humphries, J., and Slater, J. (2024). ChatGPT is bullshit. *Ethics Inf. Technol.* 26:38. doi: 10.1007/s10676-024-09775-5

Hidi, S., and Renninger, K. (2006). The four-phase model of interest development. *Educ. Psychol.* 41, 111–127. doi: 10.1207/s15326985ep4102_4

Holmes, W., and Tuomi, I. (2022). State of the art and practice in AI in education. *Eur. J. Educ.* 57, 542–570. doi: 10.1111/ejed.12533

Hong, P., Ghosal, D., Majumder, N., Aditya, S., Mihalcea, R., and Poria, S. (2024). Stuck in the quicksand of numeracy, far from AGI summit: evaluating LLMs' mathematical competency through ontology-guided perturbations. *arXiv preprint arXiv* 2401:09395. doi: 10.48550/arXiv.2401.09395

Hu, X. (2023). "Empowering education with LLMs - the next gen interface and content generation workshop [demo]." Presented at the *international conference on artificial intelligence in education* (Tokyo, Japan, July 03-07, 2023). AIED 2023.

Huggins-Manley, A. C., Booth, B. M., and D'Mello, S. K. (2022). Toward argument-based fairness with an application to AI-enhanced educational assessments. *J. Educ. Meas.* 59, 362–388. doi: 10.1111/jedm.12334

Jiang, Y., Hao, J., Fauss, M., and Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: is there a bias against non-native English speakers? *Comput. Educ.* 217:105070a. doi: 10.1016/j.compedu.2024.105070

Johnson, M. (2024). ETS principles for responsible use of AI in assessments. ETS Highlights. Available at: https://www.ets.org/Rebrand/pdf/ETS_Convening_executive_summary_for_the_AI_Guidelines.pdf (Accessed July 7, 2024).

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D, et al. (2024). Towards responsible development of generative AI for education: an evaluation-driven approach. Retrieved from: https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf (Accessed September 2, 2024).

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intell.* 6, 383–392. doi: 10.1038/s42256-024-00820-y

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., and Pineau, J. (2020). "Automated personalized feedback improves learning gains in an intelligent tutoring system." *Artificial intelligence in education: 21st international conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*, 12164, 140–146.

Lehman, B., Sparks, J. R., Zapata-Rivera, D., Steinberg, J., and Forsyth, C. (2024). A culturally enhanced framework of caring assessments for diverse learners. *Pract. Assess. Res. Eval.* 29. doi: 10.7275/pare.2102

López, A. A. (2023). Examining how Spanish-speaking English language learners use their linguistic resources and language modes in a dual language mathematics assessment task. *J. Latinos Educ.* 22, 198–210. doi: 10.1080/15348431.2020.1731693

Ma, W., Adesope, O. O., Nesbit, J. C., and Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* 106, 901–918. doi: 10.1037/a0037123

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., and Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv* 2308:02439. doi: 10.48550/arXiv.2308.02439

Mazzullo, E., Bulut, O., Wongvorachan, T., and Tan, B. (2023). Learning analytics in the era of large language models. *Analytics* 2, 877–898. doi: 10.3390/analytics2040046

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Comput. Educ.: Artificial Intelligence* 6:100199. doi: 10.1016/j.caeai.2023.100199

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Report Series* 2003, 1–29. doi: 10.1002/j.2333-8504.2003.tb01908.x

National Center for Education Statistics. (2024a). English learners in public schools. Condition of education. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/coe/indicator/cgf (Accessed July 7, 2024).

National Center for Education Statistics. (2024b). Students with disabilities. Condition of education. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/coe/indicator/cgg (Accessed July 7, 2024).

Nguyen, H. A., Stec, H., Hou, X., Di, S., and McLaren, B. M. (2023). "Evaluating ChatGPT's decimal skills and feedback generation in a digital learning game" in *Responsive and Sustainable Educational Futures. EC-TEL 2023*. eds. O. Viberg, I. Jivet, P. Muñoz-Merino, M. Perifanou and T. Papathoma (Cham, Switzerland: Springer Nature), 278–293.

Olney, A. (2023). "Generating multiple choice questions from a textbook: LLMs match human performance on most metrics." Paper presented at *the empowering education with LLMs - the next gen interface and content generation workshop at AIED 2023* (Tokyo, Japan, July 03–07).

Pankiewicz, M., and Baker, R. S. (2023). Large language models (GPT) for automating feedback on programming assignments. *arXiv preprint arXiv* 2307:00150. doi: 10.48550/arXiv.2307.00150

Pardos, Z. A., and Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS One* 19:e0304013. doi: 10.1371/journal.pone.0304013

Park, Y. (2024). "Digital-first content development for test-taker delight and fairness." Paper presented at *the 2024 annual meeting of the National Council on measurement in education*. Philadelphia, PA.

Penuel, W. R. (2019). "Co-design as infrastructuring with attention to power: building collective capacity for equitable teaching and learning through design-based implementation research" in Collaborative curriculum Design for Sustainable Innovation and Teacher Learning. eds. J. Pieters, J. Voogt and N. P. Roblin (Cham, Switzerland: springerOpen), 387–401.

Rai, P., Sood, S., Madisetti, V. K., and Bahga, A. (2024). Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on llms. *J. Softw. Eng. Appl.* 17, 43–68. doi: 10.4236/jsea.2024.171003

Randall, J., Slomp, D., Poe, M., and Oliveri, M. E. (2022). Disrupting white supremacy in assessment: toward a justice-oriented, antiracist validity framework. *Educ. Assess.* 27, 170–178. doi: 10.1080/10627197.2022.2042682

Schleifer, A. G., Klebanov, B. B., Ariely, M., and Alexandron, G. (2024). Anna Karenina strikes again: pre-trained LLM embeddings may favor high-performing learners. *arXiv preprint arXiv* 2406:06599. doi: 10.48550/arXiv.2406.06599

Sinharay, S., and Johnson, M. S. (2024). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. *J. Educ. Meas.* 61, 5–46. doi: 10.1111/jedm.12381

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educ. Meas. Issues Pract.* 39, 100–105. doi: 10.1111/emip.12377

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L. (2023). Cognitive architectures for language agents. *arXiv preprint arXiv* 2309:02427. doi: 10.48550/arXiv.2309.02427

Sun, R. (2024). Can a cognitive architecture fundamentally enhance LLMs? Or vice versa? *arXiv preprint arXiv* 2401:10444. doi: 10.48550/arXiv.2401.10444

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., et al. (2022). Assessment in the age of artificial intelligence. *Comput. Educ.: Artificial Intelligence* 3:100075. doi: 10.1016/j.caeai.2022.100075

UNESCO, IRCAI (2024). Challenging Systematic Prejudices: An Investigation into Gender Bias in Large Language Models. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000388971 (Accessed July 7, 2024).

von Davier, M. (2023). "Training Optimus prime, MD: a case study of automated item generation using artificial intelligence–from fine-tuned GPT2 to GPT3 and beyond" in *Advancing natural language processing in educational assessment*. eds. V. Yaneva and M. von Davier (New York, NY: Routledge), 90–106.

von Davier, A. A., DiCerbo, K., and Verhagen, J. (2021). "Computational psychometrics: A framework for estimating learners' knowledge, skills and abilities from learning and assessments systems," in *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*. eds. A. A. von Davier, R. J. Mislevy, and J. Hao (Springer), 25–43.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: the impact of relevant contexts on performance and learning outcomes. *J. Educ. Psychol.* 105, 932–945. doi: 10.1037/a0031882

Walkington, C., and Bernacki, M. L. (2018). Personalization of instruction: design dimensions and implications for cognition. *J. Exp. Educ.* 86, 50–68. doi: 10.1080/00220973.2017.1380590

Walkington, C., and Bernacki, M. L. (2019). Personalizing algebra to students' individual interests in an intelligent tutoring system: how moderators of impact. *J. Artif. Intell. Educ.* 29, 58–88. doi: 10.1007/s40593-018-0168-1

Wang, D., and Han, H. (2021). Applying learning analytics dashboards based on process-oriented feedback to improve students' learning effectiveness. *J. Comput. Assist. Learn.* 37, 487–499. doi: 10.1111/jcal.12502

Wang, Z., Valdez, J., Basu Mallick, D., and Baraniuk, R. G. (2022). "Towards human-like educational question generation with large language models" in International conference on artificial intelligence in education. eds. M. M. Rodrigo, N. Matsuda, A. I. Cristea and V. Dimitrova (Cham: springer International Publishing), 153–166.

Williamson, B., and Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learn. Media Technol.* 45, 223–235. doi: 10.1080/17439884.2020.1798995

Wise, S. L. (2017). Rapid-guessing behavior: its identification, interpretation, and implications. *Educ. Meas. Issues Pract.* 36, 52–61. doi: 10.1111/emip.12165

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Xhakaj, F., Aleven, V., and McLaren, B. M. (2017). "Effects of a dashboard for an intelligent tutoring system on teacher knowledge, lesson plans and class sessions." In *Artificial intelligence in education: 18th international conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18* (pp. 582–585). Springer International Publishing.

Yang, X. (2024). Linguistically responsive formative assessment for emergent bilinguals: exploration of an elementary teacher's practice in a math classroom. *Int. Multilingual Res. J.* 1–24, 1–24. doi: 10.1080/19313152.2024.2339757

Ye, H., Liu, T., Zhang, A., Hua, W., and Jia, W. (2023). Cognitive mirage: a review of hallucinations in large language models. *arXiv preprint arXiv* 2309:06794. doi: 10.48550/arXiv.2309.06794

Zapata-Rivera, D. (2012). "Adaptive score reports," in *Proceedings of the user modeling, adaptation, and personalization conference*. eds. J. Masthoff, B. Mobasher, M. Desmarais, and kambou (Berlin/Heidelberg: Springer), 340–345.

Zapata-Rivera, D., Forsyth, C. M., Graf, A., and Jiang, Y. (2024). "Designing and evaluating evidence-centered-design-based conversations for assessment with LLMs." *Proceedings of EDM 2024 workshop: Leveraging large language models for next generation educational technologies*.

Zapata-Rivera, J. D., and Greer, J. (2002). Exploring various guidance mechanisms to support interaction with inspectable learner models. *Proceed. Intell. Tutoring Syst. ITS* 2363, 442–452. doi: 10.1007/3-540-47987-2_47

Zapata-Rivera, D., Lehman, B., and Sparks, J. R. (2020). "Learner modeling in the context of caring assessments." In *Adaptive instructional systems: Second international conference, AIS 2020, held as part of the 22nd HCI international conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22* (pp. 422–431). Springer International Publishing.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2024). Explainability for large language models: a survey. *ACM Trans. Intell. Syst. Technol.* 15, 1–38. doi: 10.1145/3639372

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., and Srinivasan, K. (2024). Bias in generative AI. *arXiv preprint arXiv* 2403:02726. doi: 10.48550/arXiv.2403.02726