CASE STUDY

Efficacy, Validity and Fairness Considerations in AI-Driven Assessments

Kadriye Ercikan

Efficacy, Validity and Fairness Considerations in AI-Driven Assessments

Kadriye Ercikan

Introduction

The rapid advancements and integration of artificial intelligence (AI) are transforming the field of assessment and the science of measurement in unprecedented ways. Particularly in the last two years, generative AI has accelerated significant changes in content development, scoring, interactivity, and personalization (Arslan et al., 2024; Bulut et al., 2024; Kyllonen E. et al., 2024; Mao, Chen & Liu, 2024; Zhai & Krajcik, 2024). These AI-driven applications offer opportunities to better align assessments with educational goals by providing opportunities for:

1. Measuring complex competencies essential for success in technologically advanced educational and workplace contexts;
2. Creating engaging, interactive learning and assessment environments that are rewarding for individuals; and
3. Providing targeted feedback to support teaching and learning.

While AI can enhance assessment practices, these new approaches necessitate rigorous evaluation of their efficacy and impact on the validity and fairness of the resulting claims. This piece begins by highlighting key opportunities to innovate assessments using AI, followed by illustrative examples. The final section focuses on the critical need for evaluating the efficacy, validity and fairness of AI applications, offering guiding questions for each context.

Opportunities for Innovation in Assessment Using AI

AI offers a range of possibilities to improve assessment quality and efficiency, particularly in measuring complex constructs that have been traditionally difficult to assess through paper-based or linear digital formats (Bennett, 2024; Kyllonen et al., 2024). Notable innovations include:

- **Personalization, interactivity, and adaptivity** to engage test takers more effectively, and optimize performance on assessment;
- **Use of process data** for assessing cognitive processes; and
- **Automation of content creation, scoring, and feedback** to enhance scalability and cost-efficiency.

Below are three examples that illustrate how AI is being used to support learning, personalize assessment experiences, and increase operational efficiency.

1 Assessment to Support Learning

There is growing interest in using assessments to directly support learning. In a language learning context, AI enables learners to develop their language skills through simulations of workplace tasks. These tasks assess both language proficiency and workplace competencies and are personalized in real time based on learner choices and performance. AI provides immediate feedback and recommendations for further practice (ETS, 2024). Such AI-driven embedded assessment in authentic, real-life learning contexts offers potential for broader educational applications.

2 Personalizing Assessments

Personalization is a significant AI-enabled opportunity, with the potential to advance fairness and to meet the needs of (neuro)diverse students at scale. By tailoring assessments to align with individuals' linguistic, cultural, and educational contexts—and giving them agency in task selection—personalization provides opportunities to optimize engagement and performance (Arslan et al., 2024; Bennett, 2024). A compelling example is "Context AI," developed by Burcu Arslan and colleagues (Arslan et al., 2024). This tool uses GPT-4 to customize assessment contexts based on student's interests. For example, for a Context AI math item, the student would be allowed to pick an interest area from choices that include things like football, popular music, or gaming with Roblox. A student selecting Roblox sees an item embedded in the context of this game whereas a student picking other interest areas would have items related to those areas. Prior research shows that the context of test items can significantly influence performance (Ercikan & Solano-Flores, in press) and thus giving students the option to choose the context may be a promising direction.

> ***AI can tailor assessments to students' linguistic, cultural, and personal contexts—making them more engaging and equitable***

3 Increasing Efficiency: Human-AI Collaboration in Scoring

Earliest applications of AI in assessment has been in creating efficiencies in operational large-scale assessments (Williamson et al., 2012) and the most widely used applications of AI currently is in scoring. AI can be used to increase the efficiency of work done by humans in scoring is in two different ways. One is by replacing some of the work done by humans, and the other is to provide tools that help humans to be more efficient and improve the quality of their work. The human-AI collaboration can enhance efficiency through three complementary ways:

1. Verification
2. Contributory scoring
3. Divide and conquer

1. **Verification**: AI scores can serve a confirmatory role for human scores by confirming the human scores, and flagging disagreements. They can support quality of human scores by flagging unscorable responses. The flag and review role can also be used to verify AI scores by humans, especially in verifying scoring of “unscorable” responses.

2. **Contributory Scoring**: The scores from human raters and AI can be combined in contributory ways where human and AI scores both contribute to the final score. This approach involves both human and AI scores of the same section of the assessment contributing to the reported score. Typically, human and AI both provide holistic scores, with the automated scoring system serving the same role as a second human rater.

3. **Divide and conquer**: Humans and AI produce different kinds of information. AI might be used to evaluate specific concrete features of responses, or measure more fine-grained phenomena while humans evaluate broader and more abstract features of meaning and effectiveness. For example, in assessing and providing feedback on writing, AI could be used to provide quantitative feedback, while humans provide qualitative feedback. AI generated feedback can include a summary of the writing, evaluation of cohesion and coherence, language use such as vocabulary as well as mechanics such as spelling. Qualitative feedback might take the form of comments or coaching. In classroom contexts, teachers then can use the AI generated evaluation to personalize and provide feedback to the student. The idea is that teachers can filter the AI feedback and provide nuance around content, both saving time and allowing the teacher to focus on the things that humans do well–interpreting meaning. This kind of efficiency can be useful for teachers as well as for students who can receive the feedback in a more speedy way, than if the teacher had to read and evaluate each essay before they can provide feedback to students about their writing.

Efficacy, Validity and Fairness Considerations in AI-Driven Assessment

The transformative potential of AI in assessment demands thorough evaluations of **efficacy**, **validity** and **fairness** of claims made based on these assessments. Efficacy refers to whether the application of AI meets the intended goals. Validity is defined as the degree to which claims made based on assessments can be supported by evidence and rationales (Kane, 2006). Fairness refers to the degree to which goals are met across groups and score meaning is consistent for groups (Kane, 2012). Further specifics of these questions are highlighted below for each of the key applications.

Evaluating AI Feedback and Learning Support

When AI is used to provide automated feedback and support learning there are many factors that can influence the impact of the feedback including the nature of the feedback as well as its alignment with the individual learner. Addressing the following questions is central to evaluating efficacy, validity and fairness of this AI application:

- Does the feedback enhance engagement and learning?
- What evidence supports claims of improved engagement (e.g., process or self-report data)?
- Is there evidence of learning based on other assessments?

Is the evidence supporting improvement in engagement and learning similarly strong for individuals from different backgrounds and contexts?

Evaluating Personalization

Personalization in assessment needs to be evaluated with respect to the degree to which personalization meets its primary goals of enhancing engagement and optimizing performance (Arslan et al., 2024; Bennett, 2024). In addition, when individuals take different forms and formats of assessments in personalized assessment contexts, supporting validity and fairness of assessment claims from these assessments become a key challenge (Sinharay et al., 2025). Empirical evidence is needed to determine whether personalization:

- Increases engagement, performance and measurement precision;
- Improves alignment with learners' interests and backgrounds;
- Supports claims about personalization that can be supported by empirical evidence and rationales; and
- Enhances engagement, performance, and measurement precision consistently for individuals from different backgrounds and contexts.

The true measure of AI-generated feedback is whether it deepens engagement and learning for all students

Evaluating AI-Scoring

Automation for scoring can possibly have the greatest impact on the validity and fairness of claims from assessments. For AI based scoring a thorough evaluation needs to involve how the quality of AI based scores hold up against validity and fairness criteria, as is done for assessments that use human scores. For evaluating fairness of AI based scores we need to evaluate if the validity evidence is consistent across individuals from different contexts.

For **AI-based scoring**, key set of questions for evaluating efficacy, validity and fairness are:

- To what extent are scores consistent with human scoring?
- Is there evidence of construct irrelevant variance?
- Is there evidence of construct under representation?
- Is there evidence of systematic differences between human-AI scoring across groups from different backgrounds and contexts?

Two widely used statistics for evaluating AI-human score agreement are Quadratic Weighted Kappa and Proportion Reduction in Mean Squared Error (PRMSE). High PRMSE values (≥ 0.95) suggest AI scores can be used interchangeably with human ratings. Values ≤ 0.70, however, offer limited validity support for high-stakes use of AI generated scores (McCaffrey et al., 2022).

Concordance between human and machine scores is a requirement, however it is not sufficient for validity of claims based on scores. In addition to human-AI concordance, construct comparability of machine scores and human scores need to be evaluated. Especially given the black-box nature of AI scoring algorithms, we need to evaluate both construct irrelevant variance as well as construct underrepresentation. A special focus on construct underrepresentation is needed to ascertain specific construct components are captured by AI scoring. There are special concerns that the use of AI scoring may contribute to bias. Evaluation of fairness of AI scores need to be considered from the very beginning of the development of AI algorithms (Bennett 2024; Johnson & McCaffrey, 2023). And checks for fairness must be part of the development of AI scoring models and evaluation. This includes:

- Building scoring models using data that represent targeted populations
- Producing evidence that support fair interpretation of scores, as we do with scores from human scoring
- No subgroup difference in distribution of errors from AI scores in comparison with the human scores and
- No differential prediction of human scores (McCaffrey et al., 2022).

Final Note

AI provides many opportunities for assessments to better serve their intended goals such as providing feedback and support for learning, optimization of engagement and performance and increasing efficiency. However, these opportunities present possibilities of significant problems for assessment. Without thorough evaluations, integration of AI in assessment holds possibilities of stereotypical representation of cultural groups, inappropriate and useless feedback, inaccurate scores, and narrowing down of the measurement of the construct being targeted by the assessment. Such risks not only can harm the role of assessment in education but can have important societal impacts. The key questions I presented for each application are intended to highlight the necessity of empirical research with a variety of data sources for each application.

Group Contributions
Complex Problem-Solving Strategies
Assessing 21st Century Skills
Evaluating skills traditional tests often miss.

References

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence, 7,* 1460651.

Bennett, R. E. (2024). Personalizing Assessment: Dream or Nightmare?. *Educational Measurement: Issues and Practice, 43*(4), 119–125.

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., & Morilova, P. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges.* arXiv preprint arXiv:2406.18900.

Ercikan, K., & Solano-Flores, G. (in press). *Socio-cultural context of assessment. In Educational Measurement,* 5th Edition, Linda Cook and Mary Pitoniak, Eds.

ETS. (2024). *Converse Workplace* [mobile app]. Google Play. https://play.google.com/store/apps/details?id=org.ets.convo&utm_source=na_Med

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment.* Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education; Praeger.

Kane, M. T. (2012, March 28–29). *Validity, fairness, and testing* [Paper presentation]. Educational Assessment, Accountability, and Equity: Conference on Conversations on Validity Around the World. Teachers College, New York, NY, United States. https://www.tc.columbia.edu/media/media-library-2018/centers-amp-labs/aeri/ conferences-and-forms-/conversations-on-validity-2012-/a283745b-70b8-486c- bdee-4e1f4cda2c48.pdf

Kyllonen, P. C., Sevak, A., Ober, T., Choi, I., Sparks, J., & Fistein, D. (2024). *Charting the Future of Assessments.* ETS Research Report Series, 2024 (1), 1–62.

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends, 68*(1), 58–66.

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1), 1–58.

Sinharay, S., Bennett, R. E., Kane, M., & Sparks, J. R. (2025). Validation for Personalized Assessments: A Threats-to-Validity Approach. J*ournal of Educational Measurement.*

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice, 31*(1), 2–13.

Zhai, X., & Krajcik, J. (Eds.). (2024). *Uses of artificial intelligence in STEM education.* Oxford University Press.

About the authors

Dr. Kadriye Ercikan is the Senior Vice President of Global Research at the Educational Testing Service (ETS), President and CEO of ETS Canada Inc., and Professor Emerita at the University of British Columbia. In these leadership roles, she directs foundational and applied research. Her research focuses on validity and fairness issues and sociocultural context of assessment. Her recent research includes validity and fairness issues in innovative digital assessments, including using response process data, AI applications, and adaptivity. Ercikan is the President and a Fellow of the International Academy of Education (IAE), President of the International Test Commission (ITC), and President-Elect of the National Council on Measurement in Education (NCME). Her research has resulted in six books, four special issues of refereed journals and over 150 publications. She was awarded the AERA Division D Significant Contributions to Educational Measurement and Research Methodology recognition for another co-edited volume, Generalizing from Educational Research: Beyond Qualitative and Quantitative Polarization, and received an Early Career Award from the University of British Columbia. Ercikan is currently serving as the NCME Book Series Editor (2021-2026).

Date of Publication

November 2025

Citation

Ercikan, K. (2025). Efficacy, Validity and Fairness Considerations in AI-Driven Assessments, *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning*. University of Massachusetts Amherst Libraries