

Human–AI Collaboration in Educational Measurement: Transforming Assessment Data into Educational Action

Laura C. Egan, LauraEgan@Westat.com

Judy H. Tang, JudyTang@Westat.com

Westat

Abstract

This paper describes a framework for human–AI collaboration in educational measurement that connects four workflow components. The framework integrates educational measurement and responsible AI principles, emphasizing AI tools as support for human expertise. Illustrative applications demonstrate how human–AI collaboration can strengthen assessment processes while maintaining measurement quality and integrity.

1 Introduction

Artificial intelligence (AI) is increasingly integrated into educational measurement through applications such as process-data analytics, automated scoring, coding and classification, reporting, and feedback. Advances in machine learning, natural language processing, large language models, and agentic AI are expanding the ways assessment data can be processed, interpreted, and used. These developments offer opportunities to improve efficiency, scalability, and the educational value of survey and assessment data (Bennett, 2015; von Davier et al., 2017).

At the same time, using AI tools introduces important considerations for educational measurement. As with other advances in educational measurement, AI-supported tools and processes must continue to meet established expectations for reliability, validity, and fairness while also addressing transparency, accountability, reproducibility, and human oversight (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; National Institute of Standards and Technology [NIST], 2023; UNESCO, 2021). These considerations are particularly important when AI tools contribute to the classification,

interpretation, communication, or use of assessment evidence.

The challenge, therefore, is not simply how to apply AI tools to educational measurement, but how to integrate AI capabilities with human expertise while maintaining the quality and integrity of the measurement process. Human-centered AI emphasizes complementary roles for human judgment and computational capabilities, rather than replacing human expertise through automation (Amershi et al., 2019; Shneiderman, 2022; Yang et al., 2020).

This paper presents a *Human–AI Collaboration in Educational Measurement* framework organized around four workflow components: evidence generation, evidence processing, evidence interpretation, and educational action. The framework integrates educational measurement principles with responsible AI practices by defining appropriate roles for human expertise and AI contributions, identifying tasks that require human responsibility, and evaluating measurement across these components. Rather than beginning with AI capabilities, the framework centers the measurement process and considers where and how human experts and AI tools can work together symbiotically. Illustrative applications demonstrate how human–AI collaboration may vary across AI methods and measurement contexts.

2 Human–AI Collaboration in Educational Measurement

2.1 Integrating Educational Measurement and Responsible AI

Educational measurement provides established principles for evaluating the quality and appropriate use of assessment evidence, including reliability, validity, and fairness (AERA et al., 2014). Responsible AI frameworks incorporate considerations such as transparency, accountability, algorithmic and automation bias,

and human expert judgment and oversight (Baker et al., 2023; Goddard et al., 2012; NIST, 2023; UNESCO, 2021). The proposed framework brings these perspectives together by considering how these principles apply when AI tools contribute to generating, processing, interpreting, and using assessment evidence. Unlike broader responsible AI frameworks, it situates these principles within the evidentiary processes and requirements of educational measurement.

2.2 Centering Human Expertise

Human expertise is integral to the framework and is not limited to reviewing AI-generated outputs. Human experts define constructs and evidence requirements, establish decision rules, evaluate AI-generated outputs, resolve ambiguous cases, interpret results, and determine appropriate uses of assessment evidence. The nature and degree of human involvement may vary by AI tool type, measurement component, and the consequences of the resulting decisions.

Importantly, an AI tool's ability to perform a task does not necessarily mean the task should be delegated to it. Activities involving the interpretation of constructs, ambiguous evidence, consequential judgments, or decisions about using assessment results require human expertise and accountability. AI tools, therefore, complement, rather than replace, professional judgment.

2.3 Measurement Considerations in Human–AI Collaboration

Human–AI collaboration also requires considering how established measurement concepts apply when human experts and AI tools jointly contribute to assessment processes. These concepts do not fundamentally change; instead, AI tools introduce additional sources of evidence, variation, bias, and potential errors that may need evaluation. Reliability may include the consistency of AI-supported outputs and agreement between human and AI classifications. Validity requires evaluating whether AI-derived evidence supports intended interpretations and uses, while fairness requires examining whether AI-supported processes perform equitably across relevant populations. Accountability requires guardrails for AI's role and clarity on human responsibility for AI-supported outputs and their use.

These considerations may also differ by the type of AI tool. Machine-learning tools used for pattern

detection or classification present different measurement and oversight considerations than generative or agentic AI tools used for interpretation, feedback, or decision support. These distinctions inform the four components of the framework presented in the next section.

3 Human–AI Collaboration Framework

The *Human–AI Collaboration in Educational Measurement* framework organizes AI-supported activities across four workflow components. Human expertise spans all four components, with the nature and degree of involvement depending on the type of AI tool, the purpose of measurement, and the consequences of its use. Table 1 in Appendix A summarizes human–AI collaboration across the four educational measurement components.

3.1 Evidence Generation

Evidence generation focuses on identifying meaningful information from assessment data to understand students' knowledge, skills, behaviors, or problem-solving processes. AI tools and computational methods can identify patterns in complex data, including response processes, behavioral sequences, and other forms of process data that may be difficult to examine using conventional methods.

Human expertise is essential in determining which patterns constitute meaningful evidence and how they relate to the construct being measured. AI tools may identify associations or behavioral patterns, but measurement experts must evaluate whether those patterns provide valid and interpretable evidence for the intended measurement purpose.

3.2 Evidence Processing

Evidence processing involves transforming assessment information into forms that can be analyzed and interpreted. AI tools can support activities such as coding, classification, scoring, matching, and quality assurance, particularly when large volumes of data make exclusively manual processing resource-intensive.

Human experts establish coding and classification rules, evaluate AI-generated recommendations, review ambiguous or low-confidence cases, and conduct quality assurance. The appropriate level of AI support depends on the

task and the consequences of error. AI-supported processing therefore requires mechanisms for human review, monitoring, and adjudication rather than assuming that AI-generated classifications should be accepted without evaluation.

3.3 Evidence Interpretation

Evidence interpretation focuses on determining what assessment results mean and how to communicate them. AI tools can assist by identifying patterns, synthesizing information, generating summaries, and supporting the communication of complex assessment results.

Interpretation, however, requires consideration of the construct, context, uncertainty, and limitations of the evidence. Human experts evaluate whether AI-generated interpretations are supported by the available evidence and appropriate for the intended audience and use. Human review and adjudication are particularly important when generative AI tools produce explanations or narratives that may appear plausible even when they are not adequately supported by assessment evidence.

3.4 Educational Action

Educational action extends the interpretation of assessment evidence to feedback, instructional support, interventions, and educational decision-making. AI tools can support this component by tailoring information, generating feedback, or identifying potential recommendations based on assessment results.

The transition from evidence to action also requires careful consideration of human responsibility. AI tools may support recommendations, but decisions about how to use assessment evidence require nuanced consideration of educational context, student needs, uncertainty, and potential consequences. Human educators and other decision-makers, therefore, remain responsible for determining whether and how AI-supported information should inform educational action.

4 Human–AI Collaboration in Large-Scale Assessment

The framework applies across different AI methods and workflow components of educational measurement. Three applications from large-scale

assessments illustrate how the roles of AI and human expertise vary by measurement purpose.

4.1 Process-Data Analytics

Process-data analytics exemplifies evidence generation. Machine-learning tools can identify behavioral sequences and patterns in assessment process data that provide information beyond final responses. Human expertise is needed to determine whether the identified patterns are substantively meaningful and provide valid evidence about the knowledge, skills, or strategies under examination.

4.2 Human–Machine Learning for Qualitative Data Coding

Qualitative data coding in federal statistics uses human–machine learning to process evidence. Machine-learning tools can support coding and classify large volumes of qualitative data by identifying matches and similar responses and generating candidate classifications. Trained human coders review recommendations, resolve ambiguous or low-confidence cases, and conduct quality assurance. This example demonstrates how AI tools can reduce manual processing while retaining human expertise for coding decisions that require contextual judgment.

4.3 Agentic AI for Assessment Feedback

Agentic AI for personalized assessment feedback shows evidence interpretation and educational action. AI agents can support synthesizing and communicating student assessment information and generate individualized feedback for educators. Human expertise remains essential for evaluating the appropriateness of AI-generated interpretations and determining how assessment evidence should inform feedback or instructional decisions.

These examples demonstrate how human–AI collaboration varies across measurement contexts. The role of human expertise, use of AI tools, human responsibility, and measurement evaluations depend on the type of AI tool and its role in the measurement process.

5 Discussion and Implications

The framework extends responsible AI principles by situating them within the specific processes and requirements of educational measurement. Rather than treating or viewing

human oversight as a single safeguard, it identifies distinct roles for human expertise across evidence generation, processing, interpretation, and educational action. The appropriate form of human–AI collaboration depends on the purpose of measurement, the type of AI tool, and the potential consequences of its use.

Established measurement principles remain central when AI tools are incorporated into the measurement process, but their application may require additional considerations. Reliability may address the consistency of AI-supported outputs and agreement between human and AI classifications; validity may include evaluating whether AI-derived evidence and interpretations support their intended uses; and fairness may include examining whether AI-supported processes perform equitably across populations. Transparency and accountability are particularly important when AI tools contribute to interpretations, recommendations, or decisions.

In every part of the process, humans remain morally and legally responsible regardless of the degree to which AI agents practice operational autonomy. AI tools can enhance the scale and efficiency of educational measurement, but their ability to perform a task does not mean that the task should be delegated to them. Human expertise remains essential when tasks require knowledge of constructs, contextual interpretation, resolution of ambiguity, or consequential judgment. The use of AI tools in educational measurement should therefore be evaluated by what they can accomplish and where human expertise and responsibility must remain central to maintain the quality and integrity of the measurement process.

Acknowledgments

The authors thank the NCME AIME-Con reviewers for their thoughtful feedback and contributions to the review process and the conference organizers for their efforts in organizing and supporting the conference. Any opinions expressed here are those of the authors and do not necessarily reflect those of Westat.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and*

psychological testing. American Educational Research Association.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery.

<https://doi.org/10.1145/3290605.3300233>

Baker, R., Hawn, M.A., Lee, S. (2023). Algorithmic bias: the state of the situation and policy recommendations. In *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem* (pp. 243 – 262). OECD Publishing, Paris.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39(1), 370–407.

<https://doi.org/10.3102/0091732X14554179>

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

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

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework* (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press. <https://doi.org/10.17226/10019>

Organization for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on artificial intelligence*. OECD Publishing.

- Organization for Economic Co-operation and Development (OECD). (2019). *The survey of adult skills: Reader's companion* (3rd ed.). OECD Publishing. <https://doi.org/10.1787/f70238c7-en>
- Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press. <https://doi.org/10.1093/oso/9780192845290.001.0001>
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.
- von Davier, A. A., Zhu, M., & Kyllonen, P. C. (Eds.). (2017). *Innovative assessment of collaboration*. Springer. <https://doi.org/10.1007/978-3-319-33261-1>
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376301>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>

A Appendix

Table 1

Human-AI Collaboration Across Educational Measurement Workflow Components.

Table 1. Human-AI Collaboration Across Educational Measurement Workflow Components

<i>Measurement Workflow</i>	<i>Human Expertise</i>	<i>AI Contribution</i>	<i>Human Responsibility</i>	<i>Measurement Evaluations</i>
Evidence Generation	Define constructs and evaluate meaningful evidence.	Identify patterns and potential evidence from complex data.	Determine whether AI-derived patterns constitute appropriate evidence.	Evaluate whether AI-derived evidence is valid, reliable, and fair.
Evidence Processing	Establish rules, review outputs, resolve ambiguity, and adjudicate.	Support coding, classification, scoring, matching, and quality assurance.	Review and adjudicate uncertain or consequential outputs.	Evaluate accuracy, consistency, reproducibility, and differential performance.
Evidence Interpretation	Evaluate meaning, context, uncertainty, and appropriate interpretation.	Identify patterns, synthesize information, and generate summaries.	Determine whether AI-supported interpretations are appropriate and supported by evidence.	Evaluate whether interpretations are supported by evidence, transparent, and appropriately qualified.
Educational Action	Apply contextual and professional judgment for appropriate educational use and response.	Support feedback, recommendations, and decision support.	Determine whether and how AI-supported information should inform educational decisions.	Evaluate whether uses and decisions are appropriate, fair, transparent, and accountable.