Received: 26 March 2023 | Revised: 11 April 2023 | Accepted: 25 April 2023

DOI: 10.1002/tea.21867

COMMENT

JRST | WILEY

Can we and should we use artificial intelligence for formative assessment in science?

Tingting Li[1,2] | **Emily Reigh**[3] | **Peng He**[1,2] | **Emily Adah Miller**[4]

[1]Department of Counseling, Educational Psychology, and Special Education, College of Education, Michigan State University, East Lansing, Michigan, USA

[2]CREATE for STEM Institute, College of Education, Michigan State University, East Lansing, Michigan, USA

[3]Berkeley School of Education, University of California, Berkeley, California, USA

[4]Mary Frances Early College of Education, University of Georgia, Athens, Georgia, USA

Correspondence
Tingting Li, Department of Counseling, Educational Psychology, and Special Education, College of Education, Michigan State University, East Lansing, MI, USA.
Email: litingt1@msu.edu

In this commentary, we respond to the recent article *Applying machine learning to automatically assess scientific models* by Zhai et al. (2022). The authors present automated assessment as a solution to the problem of limited time for assessment in middle school science classrooms. Drawing from our collective expertise in science assessment, machine learning (ML), artificial intelligence (AI), and culturally relevant and linguistically responsive pedagogy, we argue that there are significant limitations to the current applications of AI for formative assessment practices. Although we believe that these limitations extend to all students, we are particularly concerned about the implications for students from nondominant cultural and linguistic backgrounds. We first share our understanding of AI's role in formative assessment, with reference to the paper by Zhai and colleagues. Next, we ask whether AI *can* effectively assess students' emergent sensemaking and then consider whether we *should* use AI for purposes of formative assessment. Finally, we discuss *how* we can better use AI for formative assessment.

1 | ARTIFICIAL INTELLIGENCE AND ASSESSMENT PRACTICES

ML is a form of AI that uses datasets to train algorithms to identify patterns and relationships, which can be used to simulate human decision-making. Zhai et al. (2021) argue that ML can

All authors have made equal and significant contributions to this paper.

advance educational assessment by capturing complex constructs, making accurate inferences from intricate data, and easing human grading efforts. In the article that is the focus of this commentary, the authors apply ML, specifically the deep learning approach, to automatically assess scientific models. This process poses a technical challenge given that models are multi-representational and include pictures and symbolic representations.

Zhai and colleagues argue that ML-based classroom assessment supports teachers in formative assessment practices, such as giving timely feedback to students and adjusting instruction. Formative assessment describes assessment that supports the process of learning (Black & Wiliam, 2009; National Research Council, 2000). By providing the teacher insight into what students know, formative assessment guides teacher decision making and promotes the development of strategies and practices for effective teaching that are tailored to a particular group of learners (Wang et al., 2010). Importantly, formative assessment should identify the range of sensemaking resources that students employ as they engage in scientific practices (Furtak et al., 2019).

2 | CAN AI EFFECTIVELY ASSESS STUDENTS' EMERGENT SENSEMAKING PRACTICES?

So, *can* AI identify students' emergent sensemaking? Zhai and colleagues argue that the multi-representational nature of models make them an equitable form of assessment. Yet, many studies raise concerns about equity in the use of AI. Cheuk (2021) argues that AI centers whiteness by taking the "culture, language, and representations of White people as the standard against which all answers ought to be seen, heard, and measured" (p. 2). The cultural and linguistic practices of students from minoritized backgrounds may not be well represented in training datasets (Yao et al., 2020). As Cheuk (2021) also points out, a major issue with the use of AI is that the processes for vetting training datasets to mitigate bias are typically absent or left undescribed. A case in point, Zhai et al. (2022) lacked background information on the students represented in the dataset, a limitation that they readily acknowledged.

Given the prevalence of bias in the training datasets, students' emergent sensemaking may not be recognized. ML algorithms tend to overvalue the language represented in rubrics, which often privilege "academic" language over students' everyday sensemaking practices (Noble et al., 2012). Using technical or specialized language can result in high scores, even if the responses lack substantial meaning (Nehm et al., 2012). In contrast, responses that convey a single concise idea, a unique perspective or use alternative wordings are often assigned low scores; responses with unexpected ideas may have no commonalities with the training dataset and are therefore often assigned a score of zero (Amerman, H., personal communication, April 6, 2023). Our work shows that human and automated scores on scientific models showed larger discrepancies for multilingual learners (MLLs) than for English-only students (Li & Adah Miller, n.d.). This result may be because MLLs' linguistic practices were not represented in the training data, and algorithms lack access to the contextual information that teachers would use to make sense of students' thinking.

Furthermore, algorithms that assess the images in models may ascribe meaning to superfluous features of students' representations. Zhai and colleagues noted that "inconsistent size" of model components may confuse computers and result in mislabeling. For instance, students who used larger or smaller arrows in their models than expected were more likely to be marked incorrect, even though this distinction did not carry meaning in the model. Across these examples, we see that algorithms are currently only able to offer valid interpretations of the

sensemaking of the subset of students who provide expected responses, which continues the cycle of reinforcement of historical inequities in whose practices are framed as "scientific."

3 | SHOULD WE RELY ON AUTOMATED FORMS OF ASSESSMENT FOR FORMATIVE PURPOSES?

Even if AI *can* eventually render more valid assessments of students' thinking, *should* we use it for formative assessment purposes? We question Zhai and colleagues' argument that automated assessment supports effective formative assessment practices. We see learning as a process of developing cultural practices through social interactions with others (Nasir et al., 2014). From this stance, we see heterogeneity as fundamental to learning (Rosebery et al., 2010) and believe that students learn best when they consider a wealth of diverse ideas (Haverly et al., 2020) and engage in expansive forms of science practices (Schwarz et al., 2022). We thus question the value of automated assessment that fits students' understandings into predetermined boxes. Rather, we think that unexpected and unique ideas may be the most useful for deepening student thinking and that recognizing these ideas is crucial to promoting equity (Rosebery et al., 2016). We underscore that using AI for formative assessment can diminish students' access to one another's sensemaking and shortchange the learning that results from creative, uncertain, and divergent thinking.

We also fear that automated scoring may undermine culturally relevant pedagogy (Ladson-Billings, 1995), which integrates students' knowledge of local places and events, cultural competence and sociopolitical consciousness into classroom activities and assessments (Mensah, 2022). The teachers in Riley and Mensah's (2023) redesigned their mandated curriculum materials "on the fly," making changes such as discussing Henrietta Lacks to elevate the sociopolitical consciousness of Black and Brown students. Critiquing science's role in upholding racist institutions is a practice that is unlikely to be recognized by an algorithm. Teachers' assessments of students are informed by their knowledge of their individual histories, group dynamics, the classroom environment, and the community context, all factors that AI cannot recognize. The proliferation of automated assessments that are used across learning contexts could draw attention away from these critical aspects of instruction.

Finally, we question the very premise that assessment needs to be faster. We argue that formative assessment is integral to teaching, rather than an activity to be outsourced. We believe the real challenge is to provide teachers with more time for formative assessment and to develop tools that help them recognize and build on the broad range of ways that students engage in sensemaking. The most unexpected and richest ideas take longer to understand, and AI, which values speed, may shortchange that process. We realize that what is ideal in teaching is not always practical, and that AI may have useful classroom applications. However, we believe that AI poses risks to the art and practice of teaching, with potential for negative consequences for student learning.

4 | WHAT IS NEXT?

We propose several lines of inquiry to potentially mitigate the limitations of AI-based assessments. Researchers should employ rigorous methodologies to validate AI-based assessments that mitigate bias for students from nondominant backgrounds (Li et al., 2022).

First, we advocate for exploration of the ways that interdisciplinary teams (e.g., educators, language specialists, community representatives, AI assessment specialists, psychologists, ethicists, and data scientists) can monitor and improve AI models (Selwyn, 2019). In the case of science, these collaborations can support the development of models that outline how to assess heterogeneity in ideas, language and practices as the expected outcome. More attention should be paid to strategizing to avoid bias (Holstein et al., 2019), which, in the case of science, requires attention to AI's treatment of student ideas that do not conform to expectations.

Next, we suggest that AI researchers explore various human-in-the-loop approaches, where teachers and automated systems work collaboratively to enhance the assessment process in ways that combine the strengths of teachers with the efficiency of AI (Holstein et al., 2019). We suggest that saving time be backgrounded as a goal, and that AI instead be used to promote teachers' interest and intellectual investment in students' sensemaking (Miller et al., 2021). We wonder how automated assessment can be leveraged to deepen and personalize teachers' interactions with students, rather than codify teacher decision making.

We applaud Zhai et al. (2022) for a window into the future of assessment, but underscore the need for more research. We must better connect theories of formative assessment, science learning, and equity with AI algorithms. Before we use AI for formative classroom use, we must find ways for it to support our ability to understand the diverse sensemaking practices that students bring to their learning. Research is needed to create and study AI models that are culturally relevant and linguistically responsive and transcend the limitations of our current systems.

ACKNOWLEDGMENTS

We would like to acknowledge Holly Amerman for comments on drafts of the manuscript.

ORCID

Tingting Li https://orcid.org/0000-0002-5692-2042
Emily Reigh https://orcid.org/0000-0003-0922-3537
Peng He https://orcid.org/0000-0002-2877-0117
Emily Adah Miller https://orcid.org/0000-0003-3473-5729

REFERENCES

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education)*, *21*, 5–31.

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education*, *105*(5), 825–836. https://doi.org/10.1002/sce.21671

Furtak, E. M., Heredia, S. C., & Morrison, D. (2019). Formative assessment in science education: Mapping a shifting terrain. In *Handbook of formative assessment in the disciplines* (pp. 97–125). Routledge.

Haverly, C., Calabrese Barton, A., Schwarz, C. V., & Braaten, M. (2020). "Making space": How novice teachers create opportunities for equitable sense-making in elementary science. *Journal of Teacher Education*, *71*(1), 63–79. https://doi.org/10.1177/0022487118800706

Holstein, K., Wortman Vaughan, J., Daumé, H., III, Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–16).

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465–491.

Li, C., Xing, W., & Leite, W. (2022). Using fair AI to predict students' math learning outcomes in an online platform. *Interactive Learning Environments*, *1–20*, 1–20. https://doi.org/10.1080/10494820.2022.2115076

Li, T., & Adah Miller, E. (n.d.). Culturally and linguistically "blind" or biased? Challenges for AI scoring of multilingual elementary students' hand-drawn scientific models. (manuscript under preparation).

Mensah, F. M. (2022). "Now, I see": Multicultural science curriculum as transformation and social action. *The Urban Review*, *54*(1), 155–181.

Miller, E. C., Severance, S., & Krajcik, J. (2021). Motivating teaching, sustaining change in practice: Design principles for teacher learning in project-based learning contexts. *Journal of Science Teacher Education*, *32*(7), 757–779. https://doi.org/10.1080/1046560X.2020.1864099

Nasir, N. S., Rosebery, A. S., Warren, B., & Lee, C. D. (2014). Learning as a cultural process: Achieving equity through diversity. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 686–706). Cambridge University Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. National Academies Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778–803. https://doi.org/10.1002/tea.21026

Riley, A. D., & Mensah, F. M. (2023). "My curriculum has no soul!": A case study of the experiences of black women science teachers working at charter schools. *Journal of Science Teacher Education*, *34*(1), 86–103.

Rosebery, A. S., Ogonowski, M., DiSchino, M., & Warren, B. (2010). "The coat traps all your body heat": Heterogeneity as fundamental to learning. *The Journal of the Learning Sciences*, *19*(3), 322–357.

Rosebery, A. S., Warren, B., & Tucker-Raymond, E. (2016). Developing interpretive power in science teaching. *Journal of Research in Science Teaching*, *53*(10), 1571–1600.

Schwarz, C. V., Ke, L., Salgado, M., & Manz, E. (2022). Beyond assessing knowledge about models and modeling: Moving toward expansive, meaningful, and equitable modeling practice. *Journal of Research in Science Teaching*, *59*, 1086–1096.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. John Wiley & Sons.

Wang, J. R., Kao, H. L., & Lin, S. W. (2010). Preservice teachers' initial conceptions about assessment of science learning: The coherence with their views of learning science. *Teaching and Teacher Education*, *26*(3), 522–529.

Yao, L., Cahill, A., & McCaffrey, D. F. (2020). The impact of training data quality on automated content scoring performance. In *The AAAI workshop on artificial intelligence for education*. Semantic Scholar. Retrieved May 7, 2023, from https://www.semanticscholar.org/paper/The-Impact-of-Training-Data-Quality-on-Automated-Yao-Cahill/8a7a0a47dd4fd989fb4431ffed2a1b3ec4b02556.

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, *59*(10), 1765–1794. https://doi.org/10.1002/tea.21773

Zhai, X., Krajcik, J., & Pellegrino, J. W. (2021). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology*, *30*, 298–312. https://doi.org/10.1007/s10956-020-09879-9

How to cite this article: Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching*, 1–5. https://doi.org/10.1002/tea.21867