Journal of Science Education and Technology (2021) 30:139–149
https://doi.org/10.1007/s10956-021-09901-8

Practices and Theories: How Can Machine Learning Assist in Innovative Assessment Practices in Science Education

Xiaoming Zhai[1]

Accepted: 15 January 2021 / Published online: 10 February 2021

Abstract

As cutting-edge technologies, such as machine learning (ML), are increasingly involved in science assessments, it is essential to conceptualize how assessment practices are innovated by technologies. To partially meet this need, this article focuses on ML-based science assessments and elaborates on how ML innovates assessment practices in science education. The article starts with an articulation of the "practice" nature of assessment both of learning and for learning, identifying four essential assessment practices: identifying learning goals, eliciting performance, interpreting observations, and decision-making and action-taking. I then extend a three-dimensional framework for innovative assessments, including construct, functionality, and automaticity, and based on which to conceptualize innovative assessments in three levels: substitute, transform, and redefine. Using the framework, I elaborate on how the 10 articles included in this special issue, Applying Machine Learning in Science Assessment: Opportunity and Challenge, advanced our knowledge of the innovations that ML brought to science assessment practices. I contend that the 10 articles exemplify a great deal of effort to transform the four components of assessment practices: ML allows assessments to target complex, diverse, and structural constructs, and thus better approaching the three-dimensional science learning goals of the Next Generation Science Standards (NGSS Lead States, 2013); ML extends the approaches used to eliciting performance and collecting evidence; ML provides a means to better interpreting observations and using evidence; ML supports immediate and complex decision-making and action-taking. I conclude this article by pushing the field to consider the underlying educational theories that are needed for innovative assessment practices and the necessities of establishing a "romance" between assessment practices and the relevant educational theories, which I contend are the prominent challenges to forward innovative and ML-based assessment practices in science education.

Keywords Machine learning · Artificial intelligence · Innovative assessment · Science

Two decades since the publication of the milestone work, Knowing What Students Know (Pellegrino et al., 2001), in which "at the heart of the committee's work was the critical importance of developing new kinds of educational assessments" (p. 1), these so-called "new kinds" of assessments continue to remain a long-standing goal for science education researchers. Many argue that the significant progress in technology, particularly in the emergent area of artificial intelligence, including machine learning (ML), may offer the potential to revolutionize assessments and meet this goal in education (Zhai et al., 2020a). For instance, a recent study (Zhai et al., 2020e) reviewed the technical, validity, and pedagogical features of ML-involved science assessments and revealed significant advantages of these innovative assessments, as compared with traditional assessments. Yet, to integrate cutting-edge technologies with assessments demands a great deal of effort to integrate assessment practices with relevant theories in education.

This article aims to forward this assessment effort specifically by focusing on applying ML to science assessment practices. I start with an articulation of the "practice" nature of assessment. This nature is by no means a novel point. The reason I attend to this point is that the field is continuingly sharing the view that assessment is purely a type of instrument, while very often ignoring the critical practices during which the participating students, teachers, researchers, and educational administrators play critical roles. I have observed this common view as an obstacle for

✉ Xiaoming Zhai
xiaoming.zhai@uga.edu

[1] Department of Mathematics and Science Education, University of Georgia, 105J Aderhold Hall, 110 Carlton Street, Athens, GA 30602, USA

pursuing assessment goals in education, especially under the guidance of A Framework for K-12 Science Education (NRC, 2012) and the Next Generation Science Standards (NGSS Lead States, 2013), and would like to further clarify this practice nature of assessment. I then attempt to conceptualize what so-called innovative assessment is. In doing so, I extend a framework developed in Zhai et al. (2020a) as a foundation for articulating innovations in assessment practices and propose three hierarchical levels. In the third section, I employ this framework and the three levels of innovative assessments to attend to the argument with regard to why and how ML has transformed science assessments. I synthesize the innovations and contributions of the articles included in the special issue, Applying Machine Learning in Science Assessment: Opportunity and Challenge (Zhai, 2019), to evidence this argument. At last, I push the field to consider the underlying educational theories that are needed for innovative assessment practices and the necessities of establishing a "romance" (the word was cited from Shavelson et al., 2008, p. 297) between assessment practices and theories, which I contend are the prominent challenges to forward innovative and ML-based assessment practices in science education. The article concludes with opportunities for future research on implementing ML in science assessment practices.

Assessment as Practices

A review of the literature in the past two decades suggests that the field is increasingly formulating a view of assessments from the perspective of "practice" (e.g., Bennett et al., 2016; Kloser et al., 2017; Zhai et al., 2018). Different from the traditional view which deemed assessments as types of educational instruments that need to be deliberately developed and implemented to support teaching, learning, and evaluation, the practice view values assessments as a series of practices of teachers, students, and other stakeholders (hereinafter I use "assessment" to indicate assessment practice and use "assessment task" to indicate the instrument of assessment). The primary goal of assessment is, through the series of practices, to collect evidence and draw valid conclusions that are informative for educational decisions both in and out of classrooms, which oftentimes result in consequential actions in education (Pellegrino, 2013).

In an era when evidence-based decisions are predominant in education and society, how educational stakeholders gather and utilize evidence to support educational decisions and actions is critical (Mercer-Mapstone & Kuchel, 2015; Shavelson et al., 2013). Assessments serve for this functionality, but they never exist in isolation—assessments are highly associated with the needs of the society and the future workforce, the national or state-wide Standards, the classroom instruction, etc. In this regard, assessments carry out two principal known purposes, assessment of learning and assessment for learning (Black & Wiliam, 1998; Wiliam, 2011). The prior highlights the functionality of assessments as a means to evaluating learning and instruction outcomes (e.g., most state-level summative assessment) and establishing accountability, or as a means of verifying the qualification and predicting future performance (e.g., professional certificate examination or college entrance examination). Therefore, the NRC Assessment Report (Pellegrino et al., 2014) refers to assessments for this type of purpose as monitoring assessment, which is prevalent in state or national level assessments. For example, the recently published Criteria for Summative Science Assessments (NGSS Lead States, 2018) that followed the NGSS (NGSS Lead States, 2013) explicitly state, "Assessments are intentionally designed to assess state science standards to provide evidence to support, refute, or qualify state-specific claims about students' achievement in science" (p. 9). To be noted, the Criteria specifically highlight "connecting evidence and use," which will involve researchers, administrators, teachers, and students and may consequently impact educational policy, students' development, and career paths, as well as the future STEM industry.

The second purpose of assessments (i.e., assessment for learning), in contrast to accountability and qualification, is primarily aimed at collecting feedback to support teachers' instructional decision making and students' learning. Assessment practices for this purpose are usually incorporated into classroom instruction and sometimes referred to as assessment as learning (Hickey et al., 2012). Depending on the primary agents for decision making, such assessments can be further differentiated as either teacher- or student-oriented. In the study about informal formative assessments, Ruiz-Primo and Furtak (2007) highlight how assessment practices support teachers in decision making. They identify an ESRU cycle, in which classroom activities are specified as teachers Elicit student thinking, Students respond to questions, teachers Recognize students' responses, and then Use the information to support students' learning. While Ruiz-Primo and Furtak's cycle outlines how assessment practices support teachers' instructional decision making, Shepard et al. (2018, p. 21) further emphasize the student-oriented classroom assessments—"[assessment] should avoid using points and grades 'to motivate' students but should create opportunities for students to use feedback to improve their work." Though paying varying degrees of attention to students as compared with teachers, both views consistently regard assessment practices as a means of collecting and utilizing evidence for facilitating learning. These points had been addressed in the seminal work of Black and Wiliam (1998) as well, in which they stated,

> We use the general term assessment to refer to all those activities undertaken by teachers—and by their students in assessing themselves—that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs. (p. 140).

Though the two types of purposes vary significantly in assessment practices, their boundary is often blurred (Bennett et al., 2016). Viewed at levels from national and state to classroom, assessments may serve both purposes simultaneously, although with varying degrees of emphasis in a specific context. There were efforts in the field to break the distinction between both purposes, proposing assessment as learning (Hickey et al., 2012). Yet, there is no wide consensus in the field and much research continues to distinguish the two types of purposes. This is evidenced by the fact that a large body of studies has articulated the practice nature of assessments for the second type of purpose (Nicolaidou et al., 2011), while relatively less attention was paid to the practice nature of assessment with the first purpose (Penfield & Lee, 2010).

Despite such existing "discrepancies," both assessment purposes could be achieved through a series of shared practices: identifying learning goals, eliciting performance, interpreting observations, and decision-making and action-taking using the assessment information.

Identifying Learning Goals. Assessment practices start from specifying the learning goals that the assessment is tapped into. Specifically in science education, the Framework for K-12 Science Education (NRC, 2012) has identified the learning goals of science as three integrated dimensions: science and engineering practices, crosscutting concepts, and disciplinary core ideas. The Next Generation Science Standards (NGSS Lead States, 2013) further specify a set of performance expectations for students in different grade bands. These learning goals should be embedded in the assessment tasks so that teachers or researchers could align assessment practices with standards, curriculum, and instruction.

Eliciting Performance. To infer students' cognition and proficiency, and justify the intended claims made on the uses of assessment scores, one needs to develop prompts or tasks to elicit students' performance. The performance, in various forms such as scientific practices or problem-solving, would provide evidence to infer the invisible state of students' minds.

Interpreting Observations. The observations of students' performance may be evaluated using rubrics that reflect the learning goals, resulting in scores. To make sense of the scores, one needs to interpret the scores to understand students' thinking and to what degree the learning goals are met.

Decision-Making and Action-Taking. Depending on the purpose of assessment, decisions made on the assessment outcomes may vary significantly. That is, a single assessment result may be used in significantly different ways, resulting in different actions and consequences (Kane, 2013).

To be noted, the series of assessment practices as a whole formulate effective assessments for both purposes mentioned above. Individual practice is not necessary to stand as assessment. Also, the proposed sequence of assessment practices is not linear and sequential. Practitioners may go back and forth across practices before the final decisions are made and the consequential actions are taken. Each of the practices is complicated and requires teachers' or researchers' essential assessment competency and pedagogical content knowledge to be successful. Moreover, assessment practices are not likely to improve without utilizing necessary technologies (e.g., DeBoer et al., 2014). Essential technologies integrated may further innovate the assessment practices and result in more informative educational decisions and actions.

Innovative Assessment[1]

The field has a long history of employing technologies to improve science assessment practices. Notable efforts were made through utilizing technologies such as drawing tools (Chang et al., 2013), mobile technology (McMahon et al., 2016), classroom adaptive learning system (Beatty & Gerace, 2009), augmented reality (Ferrer-Torregrosa et al., 2015), machine learning (Nehm et al., 2012), web-based inquiry (Liu et al., 2011), and automatic guidance (Zhai, 2021). Technologies like these extended the nature of the problems that can be presented, as well as the approaches of eliciting and interpreting evidence, and thus

[1] Author note. In the preparation of the manuscript From Substitution to Redefinition: A Framework of Machine Learning-based Science Assessment (hereby called ML Framework; Zhai et al., 2020a). I benefited from a long conversation with my co-author Mark Urban-Lurain, the co-founder of Automatic Analysis of Constructed Responses (AACR) group at Michigan State University, and Kevin Haudek, the current lead Principal Investigator of AACR. Urban-Lurain has dedicated more than 20 years in studying the use of technology in STEM education, particularly on constructed response assessments, and had valuable insights about multiple aspects of applying technology, such as machine learning (ML), in advancing STEM assessments. The conversation started from a comment Urban-Lurain made on how we should conceptualize "innovative assessments." More specifically, he urged me to think why we could possibly use the word "redefinition" to portray "innovative assessment." If assessment is still in forms of multiple-choice, constructed responses, etc., and teachers and students are involved in activities that they used to do, what do we mean by "innovative" to a degree that the assessment may be "redefined" by ML? This section provides information partially serving as a response to Urban-Lurain' question.

enhancing the assessment practices. These changes refine assessment practices and thus are regarded as "innovations." Even though innovative assessments are desired, and a considerable of innovations have been developed in the field, there are limited efforts made to conceptualize the degree to which these technologies innovate assessment practices. In this section, I first extend a framework from Zhai et al. (2020a) to conceptualize innovative assessment and then propose three levels of innovations in assessment practices.

A Conceptual Framework for Innovative Assessment

In our recent study (Zhai et al., 2020a), we developed a framework to conceptualize assessment innovations, including three dimensions: construct, functionality, and automaticity. While the framework was originally developed constrained to a specific research context and purposes, I extend it in this article to adapt it to a broad concept of innovative assessment. The three dimensions may be referred to when conceptualizing the innovations of assessment practices.

The construct is conceived as a latent trait of examinees that a test is intended to test, which accounts for the examinees' performance on the test (Cronbach & Meehl, 1955). In contemporary assessment practices, the construct sits in the center of the design of relevant tasks and the validity (Pellegrino et al., 2014). Based on this idea, a construct-centered design approach was articulated in the NRC Assessment Report (Pellegrino et al., 2014), which starts from specifying the construct or competence as a domain of knowledge or skills that the assessment task is intended to assess, which differs from the traditional task-centered design which begins with a specific activity or from which one can score particular knowledge (Messick, 1994). In science education, constructs are domain-specific and complex because of the very nature of science learning. Therefore, conceptualizing the features of a construct that science assessment tasks may tap into is critical to understand how innovative the assessments are. In our research (Zhai et al., 2020a), we have proposed three essential features for construct. The first looks at how Complex the construct is, referring to Bloom's taxonomy (Anderson et al., 2001). We argued that the higher-order complex constructs such as evaluating and creating (e.g., argumentation, modeling) are more challenging to assess, compared with those simpler ones. Therefore, assessment developers should pursue innovative approaches to tapping those complex constructs in science education. The second feature, which Zhai et al. (2020) argued that researchers should consider, is the Diversity of construct, which indicates the kinds of combination of cognitive demands that the assessment task is intended to measure. The third feature refers to the Structure of the construct, which specifically indicates the degree to which the assessment tasks may be aligned with students' cognitive development. Assessments should aim at developing a novel approach to gauging cognitive development so that we can better understand students' domain knowledge and skills in science.

Assessment practices are fundamentally a process of evidentiary reasoning, in which the functionality of assessments denotes how assessors elicit, interpret, and use the evidence to make useful decisions. To conceptualize the innovation of the practices, particularly in science education, one may consider these aspects:

Task authenticity. Assessment tasks that mimic authentic work that scientists are involved in, such as modeling and argumentation, are desired to better elicit students' knowledge, thinking, and competency. While this is not always feasible, educational technology has great potential to help us innovate towards this goal. Examples are those such as using augmented reality (e.g., Ferrer-Torregrosa et al., 2015), simulation (e.g., Frezzo et al., 2010; Gale et al., 2016), or web-based inquiry (Gobert & Pallant, 2004) to engage students in assessments.

Digital representation. One significant challenge in assessment practice is that student thinking is invisible. Coping with this challenge offers opportunities for utilizing educational technology to present and visualize students' cognitive processes so that teachers and researchers may better infer students' cognition (Jescovitch et al., 2020).

Evidence diversity. The rapid development of technology has created various new approaches to eliciting evidence of students' science learning. For example, using online discussion boards to collect evidence of the emotional and linguistic features of students could supplement traditional paper-pencil tests (Yoo & Kim, 2014). This approach provides interactive data so that teachers could better understand students' thinking processes, but it would not be accomplished without technologies. By using technologies, assessors are able to make decisions based on rich and diverse evidence, which is beyond that collected from traditional paper–pencil tests.

Measurement models. To associate scores that students earned with the levels of performance on a specific construct, measurement models may serve as tools to quantify such connections and visualize potential patterns (Wilson, 2005). Developing such methodology and tools, including computer algorithms, is a growing field as technology is increasingly involved in the processes.

Score uses. Depending on the purpose of score uses, assessment practices may present different challenges for the assessors and other stakeholders. For the state-level summative assessments, which are usually high-stakes, accuracy and equity come into the front. Technologies are often used for delivery, grading, etc., and often draw concerns. For those classroom assessments, integrating with curriculum and instruction, gauges of nuance, timely feedback may be of those that are primarily concerned with. Educational technologies that foster automatic scoring, automatic guidance, and adaptive learning have been broadly adopted toward these uses.

The third dimension looks at automaticity, which is the degree to which computers may conduct tedious and repetitive work to ease humans' effort. By improving automaticity, one could expect to increase the efficiency of assessment practice at scale (Bennett, 2018). Moreover, it is possible that by means of automatic scoring, automatic feedback, automatic guidance, etc., students and teachers may significantly benefit from these technology-enhanced assessments. In terms of automaticity, Zhai et al. (2020a) specifically emphasize the generalizability of algorithmic or statistical models for assessments. That is, to what extent a model could be applied to new scenarios and to solve new problems. Problems of this kind are still a great challenge in assessments, even with the most cutting-edge technology such as ML (Zhai et al., 2020e).

Levels of Innovative Assessment

According to the three dimensions above mentioned, innovative assessment can be conceived as assessment practices being improved through using technologies, in terms of the targeted construct, the functionality of evidence collection, interpretation, and use, as well as the automaticity of easing the human effort and supporting immediate decision-making and action-taking. To be noted, it is by no means obvious that a given technology may innovate assessment practices in isolation. Integrating technology in assessment practices is as challenging as the technological innovation itself, if not more. Thus, according to how assessment practices are innovated using technologies, the innovative assessment may be characterized as being in one of three levels: substitute, transform, and redefine.

Substitute indicates that technologies are used in assessment practices to substitute functions of conventional assessments but not improve the quality of decision-making and action-taking. For example, at the early age of using computer-embedded assessments, teachers employed computer programs to automatically score students' multiple-choice items after class. In this case, technology was used to substitute teachers' scoring work and save teachers' time for scoring, but the approaches to collecting, interpreting, and using assessment information were not changed compared with conventional assessments. It is not likely to expect a higher quality of decision-making or action-taking in this scenario, given the approach of technology use. Clark (1983) dismisses this type of innovation using a metaphor that technology acts as the "mere vehicles that deliver instruction but do not influence student achievement any more than the truck that delivers our groceries causes changes in our nutrition" (p. 445).

Transform is a higher-level innovation compared with substitute. When assessment practices are transformed by technologies, the decision-makers may obtain the opportunity to make better decisions and take actions that are more effective. Transformed assessment can be in many different forms, according to the three dimensions of the framework. In the above example, if the automatic scoring approach is used with clickers so that teachers could acquire students' scores in a timely fashion and make immediate adjustments to the next-step teaching, assessment practices may be transformed. Transformed assessments can also improve the targeted construct to be more complex, diverse, and structural. For example, the Programme for International Student Assessment recently incorporated simulations so that assessors could measure students' knowledge-in-use during practices, which might not be possible using conventional paper-pencil assessments. Such assessment practices offer innovative approaches to collecting evidence and may support assessors to better understand students' science competence. These assessments could tap into more complex constructs and support better decision making, thus are regarded as being transformed.

Redefine indicates a radical innovation of assessment practices, which is above the level of transform. According to Zhai et al. (2020a), redefine is conceptualized based on not only the constructs, functions, and automaticity but also the degree to which assessment practices could address fundamental challenges in individual learning, alignments between standards, curriculum, instruction, and assessment, accountability, administration and graduation, promotion, or developmental course placement. This level is especially difficult to achieve as a radical revolution in assessment practices would require decent integration between technology, relevant educational theories, and assessment practices. For example, to address the alignment between standards, curriculum, instruction, and assessment, the assessment tasks need to (a) be designed to elicit the constructs denoted in the standards, (b) reflect the development of student cognition, (c) be instructionally sensitive, (d) provide immediate feedback that is accurate, reliable, and interpretable, and (e) align the feedback with curriculum materials so that teachers and students could use the feedback to adjust teaching and learning immediately.

In the following section, I review the innovations of ML-based assessments in the articles included in the special issue, Applying Machine Learning in Science Assessment: Opportunity and Challenge (Zhai, 2019), using the three-dimensional conceptual framework of innovative assessment practices.

What Potential Does Machine Learning Have for Innovative Assessment Practices?

Many argue that ML may be a super "bridge" to connect the learning goals and educational decision-making which potentially could attend to the goal of redefining assessment practices (Zhai et al., 2020b). ML belongs to the family of artificial intelligence. ML uses a novel approach for automaticity which may conduct complex tasks like human beings do. That is, ML "learns" how to conduct tasks from humans, which is different from conventional computational technologies which primarily execute commands that humans enter. Through the "learning" process, computers work with data to establish algorithmic models and then validate the models (Zhai et al., 2020e). Once the models are found to have high accuracy for prediction or classification, they could be applied to predict or classify new data, such as scoring student performance on assessment tasks. The articles included in this special issue provide evidence that has significantly contributed to our understanding of how this technology might innovate assessment practices.

ML Allows Assessment Practices to Target Complex, Diverse, and Structural Constructs, and Thus Better Approaching the Science Learning Goals

One reason that ML is particularly useful in science education is due to the very nature of science learning. Since the Framework for K-12 Science Education (NRC, 2012) setting forth a new vision of science learning, the field has increasingly realized that the nature of science learning is an integration of science and engineering practices, crosscutting concepts, and disciplinary core ideas (three-dimensional learning). To meet the vision, science educators have to engage students in practices to improve students' competence to construct explanations, figure out solutions, and solve problems. The articles in this special issue made substantial contributions by tapping into science learning that is embedded with such complex scientific practices such as modeling (Zhai et al., 2020c), scientific argumentation (Lee et al., 2021; Wang et al., 2020), investigation (Maestrales et al., 2021), multimodal representational thinking (Sung et al., 2020), explanation (Jescovitch et al., 2020), and epistemic knowledge of model-based explanation (Rosenberg & Krist, 2020). For example, in their study, Maestrales et al. (2021) employed ML to automatically score students' performance by the dimension of science learning and achieved high scoring accuracy. Research both by Jescovitch et al. (2020) and Wang et al. (2020) aligned their machine scores with learning progressions, the developmental cognitive features of students' learning. These applications demonstrated the great potential of ML to tap complex, diverse, and structural constructs in science learning. Success in automatically assessing students' science learning in these practices enabled teachers and students to focus on and achieve NGSS-aligned learning goals.

ML Extends the Approaches Used to Eliciting Performance and Collecting Evidence

It is well acknowledged that the vision of engaging students in three-dimensional learning is not likely to be fully realized without decent performance-based constructed response assessments. Performance-based constructed response assessments are conceived as more efficient ways of collecting evidence to reflect students' knowledge-in-use ability (Harris et al., 2019), compared with the multiple-choice format ones (Darling-Hammond, 2014). This is because multiple-choice items are difficult to elicit higher-order thinking that is associated with sophisticated cognitions and performance. The articles in this special issue extended the approaches to collecting evidence in ways such as virtual reality (e.g., Sung et al., 2020), representations (e.g., Zhai et al., 2020c), and facial expression identification (e.g., Liaw et al., 2020). In their study, Sung et al. (2020) employed the augmented reality technology with a thermal camera attached to a smartphone to elicit students' understanding and asked students to write constructed responses; students' responses were analyzed using a deep learning approach. Zhai et al. (2020c) demonstrated how to apply deep learning to automatically evaluate students' modeling competency by scoring their drawing and writing explanations. Liaw et al. (2020) also employed a deep learning-based app to automatically identify students' emotion and examined how their emotion was associated with their learning outcomes. While these measures are only likely to be achieved with novel technology, it should also be noted that these measures have significantly extended humans' ability to infer students' competence and thinking by examining complex products developed by students or their problem-solving procedures.

ML Provides a Means to Better Interpret Observation and Use Evidence

Statistics are playing increasingly important roles in interpreting and utilizing data to infer students' cognition and predict their performance for certain work. However, statistical models have many assumptions to meet before the models can be applied to data. Especially when too many variables are involved in a model, it is difficult to meet the assumptions (e.g., multivariate normality). ML is relatively flexible on this, as the major concern for ML is the prediction accuracy (Zhai et al., 2020a). Therefore, ML may be used for a large dataset with large numbers of variables where traditional statistics would not work. Because of this feature, ML could help educators identify and interpret patterns in big datasets. For example, Bertolini et al. (2021) employed a dataset with 53 variables from 3225 undergraduate students to develop algorithms to predict students' attrition. It is anticipated that once we have such models, educators may use the models to predict students' attrition and then develop better strategies to support those students who may have a high risk of attrition. Datasets such as these may have limited use without ML in the past because it is difficult to interpret the information encapsulated; yet, they are increasingly playing significant roles in educational decision-making with the help of ML.

ML Supports Immediate and Complex Decision-Making and Action-Taking

One of the most significant contributions of ML-based assessments is timely feedback, which enables teachers and students to make instructional decisions and take actions almost immediately. The feedback could be used to provide automatic guidance for science learning. In their study, Lee et al. (2021) developed an automated feedback system that could provide students with feedback on their written arguments so that students could self-remediate their arguments. This system was designed to support students' adapted learning, which personalized students' learning experience as the feedback was individualized. Lee et al. (2021) research demonstrated the great potential that ML could support students' decision making and action taking by providing personalized feedback.

Besides the above-mentioned, the articles also make other technical contributions to the field. For example, both studies by Rosenberg and Krist (2020) and Lamb et al. (2020) explored combining ML and other statistical modeling approaches to better interpreting data and inferring students' thinking and knowledge. Both Jescovitch et al. (2020) and Wang et al. (2020) compared the analytic and holistic scoring approaches for ML scoring. They both aligned their assessment tasks with learning progressions, though associated with different content and scientific practices. Interestingly, both studies yielded consistent conclusions: machines generated more accurate results by using the analytic scoring approach as compared with using the holistic approach. Zhai et al. (2020c) specifically examined the validity issues of ML-based next generation science assessments and proposed a validity inferential network to guide assessment development and validity conclusions made using ML scores.

Could we say that ML has redefined science assessments? After reviewing and editing this special issue, I believe that we are getting closer thanks to the contributions of the articles. As compared with a recent review study (Zhai et al., 2020e), the articles in this special issue employed more matured deep learning algorithms in assessing drawing (Zhai et al., 2020c), epistemic knowledge (Rosenberg & Krist, 2020), and higher-order thinking in science (Lamb et al., 2020; Sung et al., 2020). I realized that the approaches to collecting evidence were also improved because of ML. For instance, in their study, Liaw et al. (2020) employed a facial expression identification approach to capturing students' emotional information during learning and applied ML algorithms to predict student learning outcomes. It is also delightful to find that the timely feedback assisting students in meeting challenges in scientific practices was improved by using ML (e.g., Lee et al., 2021). All the technological progress indicates that we are on the right track to redefine science assessment by employing ML.

Establishing a "Romance" Between ML-Based Assessment Practices and Theories

Effective ML-based assessment practices have to be founded on decent theories. Though the articles in this special issue have significantly advanced these emerging ML-involved assessments in many regards, there are remaining challenges that need to be addressed. For example, we are not quite sure why machine scoring capability varies significantly according to assessment internal (e.g., features of the construct) and external features (e.g., the form of the task), as well as examinee features (Zhai et al., 2020d). Research is needed to study how to better implement these advanced innovative assessment practices to promote science learning. It is essential to continue to examine theories and incorporate them into innovative assessment practices to further advance this field of research.

Domain-Specific Learning Theory

To better serve both purposes of the assessments above mentioned, innovative assessments such as those involving ML should incorporate domain-specific learning theories. In the past decades, the theories of science learning have been greatly advanced because of the development of cognitive science, epistemology, and sociocultural science (Duschl, 2008; Pellegrino, 2018). The field is accumulating knowledge and shaping a better understanding of the origins, scopes, and nature of scientific knowledge, as well as how one develops scientific knowledge (Kelly et al., 2012). Seminal research such as those revealed the differences between scientists' and novice' learning and problem-solving (e.g., Chi et al., 1981), uncovered how one develops models to explain phenomena (Clement, 2000), improved scientific literacy during investigation (Abd-El-Khalick et al., 2004), and formulated ideas in argumentation (Osborne, 2010) has provided a foundation for complex assessment practices. To portray students' developmental features, research in the past more than 10 years has developed learning progressions for big ideas and core competence of science (e.g., Osborne et al., 2016; Schwarz et al., 2009). These prominent advances, accompanied by the learning goals presented in the NGSS (NGSS Lead States, 2013), need to be incorporated into innovative assessment practices.

Validity Theory and Assessment Design Principles

Given that the involvement of ML may affect many aspects of assessment practices, it is essential to re-examine the validity theory that is commonly applied in conventional assessments. For instance, ML has significantly advanced the learning goals that assessments may tap into and transformed the approaches of collecting evidence, as well as interpreting the evidence. These changes are likely to draw new validity risks that conventional assessments never attended to. For example, new construct-irrelevant variances may appear in student scores because of the involvement of technology for the delivery of tests, collecting evidence, and interpreting scores (Zhai et al., 2020b). Computer algorithms trained using biased data may generate biased outcomes (e.g., Google Photos tags African Americans as gorillas through facial recognition; Zhang, 2015) and thus result in biased decisions and actions in education. Potentially, these biased decisions and actions potentially can lead to ethical and equity problems in education (Zhai et al., 2020c). Without considering these risks, one may draw invalid conclusions from the assessment results (Mislevy, 2016). In this regard, Zhai et al. (2020c) have provided a framework to account for the cognitive, instructional, and inferential validity of ML-based science assessments. After examining the major validity issues of ML-based assessments, they laid out the major claims, assumptions, and inferences, as well as the potential validity evidence. The framework could serve as a foundation for developing and validating future innovative assessments that are aligned with the NGSS (NGSS Lead States, 2013).

At the same time, I encourage science educators to employ either the evidence-based design (Mislevy & Haertel, 2006) or the construct modeling approach (Wilson, 2005) suggested in NRC Assessment Report (Pellegrino et al., 2014) for task development.

Technology Integration Theory

Technology plays a critical role in innovative assessment practices, but technology itself does not stand in isolation. It is evidenced that the articles in this special issue present examples of how technologies (e.g., ML, simulation, facial expression identification) are integrated with other components of assessment practices. These technologies support the development of authentic assessment tasks (Sung et al., 2020), collecting data (Liaw et al., 2020), automatically scoring student responses (Rosenberg & Krist, 2020), etc. However, it should be noted that technology can be useful only if it is situated within authentic task scenarios and integrated with appropriate learning goals (Neumann & Waight, 2020). The development of effective assessment tasks should also follow principles so that the technology could serve the purposes of the assessment practices. Krajcik and Mun (2014) recommended five principles to integrate technologies in science learning, among which four should be considered for assessment integration: (a) situate assessment tasks in authentic and real-world contexts, (b) use technology as a cognitive tool to elicit students' performance that may serve as evidence to infer their cognition, (c) align technology with the science learning goals, and (d) provide appropriate scaffolding in assessment practices using technology (e.g., Lee et al., 2021). Successful implementation of ML-based assessments should consider how to apply these principles to better integrate all technologies involved in the assessment practices.

Pedagogical Theory with Assessment Practices

Innovative assessment is a critical component of pedagogical practices, during which teachers use the assessment practices to engage students in knowledge-in-use science learning. Research has indicated that teachers usually lack the necessary pedagogical content knowledge to engage students in effective assessment practices, particularly with regard to the NGSS-aligned innovative assessments (Harris et al., 2019). Given this fact, it can be anticipated that teachers would face even more challenges when using ML-based assessments. This is because ML enables innovative assessment tasks to be used in classrooms, but teachers have no such assessment experience. Even though ML may provide means for teachers to access automatic scores of students' performance on complex science learning, teachers have to use the information to make instructional decisions by themselves. What makes it challenging is that most of the decisions have to be made immediately in classrooms, which need essential pedagogical content knowledge (Magnusson et al., 1999). In the case when automatic feedback is available for students' scientific practices, teachers may need to consider and adjust their roles in teaching and identify approaches to better supporting students (Zhai, 2021). Consistent with the findings in the review study (Zhai et al., 2020e), a review of this special issue suggests that more studies toward implementing innovative assessment practices are needed. Empirical evidence is needed to understand how to effectively incorporate pedagogical theories with innovative assessment practices.

Remark

Similar to other cutting-edge technologies, ML is ambitiously positioned in science education to not only substitute human scoring but also redefine the fundamental nature of science assessment practices (Zhai et al., 2020a). To assist this effort, this special issue called for studies to explore opportunities to redefine traditional assessment practices by means of (a) assessing complex constructs that are difficult to capture (e.g., three-dimensional science learning) with traditional assessment tasks; (b) improving the approaches we use to collecting evidence and inferring students' cognition, as well as enriching the types of evidence we use to make decisions (e.g., psychological data by sensors); and (c) advancing the automaticity of the process by easing the demands of human work on assessment. It is delightful to find that the ten articles presented in this special issue exemplify a great deal of effort toward this goal. As discussed in this article, I have found that these articles provided insights on the development and use of such ML-based assessment tasks, as well as engaging students and teachers in such assessment practices. These articles also represented efforts to meet the vision of the Framework for K-12 science education (NRC, 2012) and the NGSS (NGSS Lead States, 2013). ML-based assessments potentially could meet assessment challenges because of their ability to tap complex constructs, extend approaches to collecting evidence and inferring students' scientific thinking, and provide feedback to support timely decision making. These innovative assessment practices are essential to align the Standards with curriculum, instruction, and assessments, and are aligned with the call in the NRC Assessment Report (Pellegrino et al., 2014),

> Existing and emerging technologies will be critical tools for creating a science assessment system that meets the goals of the framework and the NGSS, particularly those that permit the assessment of three-dimensional knowledge, as well as the streamlining of assessment administration and scoring. (p. 8).

However, we are still at the transformation stage of applying ML and other artificial intelligence technologies in supporting educational decision-making and action-taking in science education. Many problems such as those regarding validity, equity, and pedagogy need to be studied to better serve the educational goals in science. The fact that ML "has transformed yet not redefined assessment practices" calls for establishing a "romance" between practices and relevant theories. As more applications of ML appear in science assessment practices, the long-standing goal of redefining science assessment practices is in the near future to come.

Acknowledgements The author is grateful to Mark Urban-Lurain for his feedback and comments on this manuscript. A special thanks and appreciation to the commentator (Joseph Krajcik) and the reviewers of the special issue. The reviewers for this special issue are as follows: Alex Lyford, André Rupp, Deborah Harris, Kevin Moore, Kevin Raczynski, Klint Kanopka, Knut Neumann, Lehong Shi, Shiyan Jiang, Tina Cheuk, Joshua Rosenberg, Len Annetta, Ronald Persin, Fabian Zehner, Ross H. Nehm, Georgia Wood Hodges, Hee-Sun Lee, Kihyun Ryoo, Christopher Harris, Jari Lavonen, Brian Gane, Peng He, Kevin Haudek, Jeanne Kriek, Megan Shiroda, Mark Urban-Lurain, Ji Shen, Mei-Hung Chiu, Yoav Cohen, Christopher Nakamura, Hsin-Yi Chang, Jonathan Osborne, Dirk Krüger, Srilakshmi Anireddy, Kamali Sripathi, Xiufeng Liu, John Stewart, Amy Pallant, Janice Gobert, and Susan Lottridge

Funding The work is partially supported by the National Science Foundation under Grant 1561150.

Declarations

Disclaimer The ideas presented in this article belong to the author and do not represent the view of the National Science Foundation.

Conflict Interest The author declares that he has no conflict of interest.

Ethics Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Informed Consent The author agreed to publish the study in this journal.

References

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., & Hofstein, A. (2004). Inquiry in science education: international perspectives. *Science Education, 88*(3), 397–419.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Longman

Beatty, I. D., & Gerace, W. J. (2009). Technology-enhanced formative assessment: a research-based pedagogy for teaching science with classroom response technology. *Journal of Science Education and Technology, 18*(2), 146–162. https://doi.org/10.1007/s10956-008-9140-4.

Bennett, R. E. (2018). Educational assessment: what to watch in a rapidly changing world. *Educational Measurement: Issues and Practice, 37*(4), 7–15

Bennett, R. E., Deane, P., & van Rijn, W. P. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist, 51*(1), 82–107.

Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Testing the impact of novel assessment sources and machine learning methods on predictive outcome modeling in undergraduate biology. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09888-8.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Chang, H.-Y., Quintana, C., & Krajcik, J. (2013). Using drawing technology to assess students' visualizations of chemical reaction processes. *Journal of Science Education and Technology, 23*(3), 355–369. https://doi.org/10.1007/s10956-013-9468-2.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459

Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education, 22*(9), 1041–1053

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 1–28

Darling-Hammond, L. (2014). *Next generation assessment: Moving beyond the bubble test to support 21st century learning*. John Wiley & Sons

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., & Buckley, B. C. (2014). Comparing three online testing modalities: using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching, 51*(4), 523–554.

Duschl, R. (2008). Science education in three-part harmony: balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*(1), 268–291

Ferrer-Torregrosa, J., Torralba, J., Jimenez, M. A., García, S., & Barcia, J. M. (2015). ARBOOK: development and assessment of a tool based on augmented reality for anatomy. *Journal of Science Education and Technology, 24*(1), 119–124. https://doi.org/10.1007/s10956-014-9526-4.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2010). Design patterns for learning and assessment: facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology, 19*(2), 105–114. https://doi.org/10.1007/s10956-009-9192-0.

Gale, J., Wind, S., Koval, J., Dagosta, J., Ryan, M., & Usselman, M. (2016). Simulation-based performance assessment: an innovative approach to exploring understanding of physical science concepts. *International Journal of Science Education, 38*(14), 2284–2302. https://doi.org/10.1080/09500693.2016.1236298.

Gobert, J. D., & Pallant, A. (2004). Fostering students' epistemologies of models via authentic model-based tasks. *Journal of Science Education and Technology, 13*(1), 7–22

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67. https://doi.org/10.1111/emip.12253.

Hickey, D. T., Taasoobshirazi, G., & Cross, D. (2012). Assessment as learning: enhancing discourse, understanding, and achievement in innovative science curricula. *Journal of Research in Science Teaching, 49*(10), 1240–1270.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2020). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, 1–18

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73

Kelly, G. J., McDonald, S., & Wickman, P. O. (2012). Science learning and epistemology. In *Second international handbook of science education* (pp. 281–291). Springer

Kloser, M., Borko, H., Martinez, J. F., Stecher, B., & Luskin, R. (2017). Evidence of middle school science assessment practice from classroom-based portfolios. *Science Education, 101*(2), 209–231. https://doi.org/10.1002/sce.21256.

Krajcik, J. S., & Mun, K. (2014). Promises and challenges of using learning technologies to promote student learning of science. *Handbook of Research on Science Education, 2,* 337–360

Lamb, R., Hand, B., & Kavner, A. (2020). Computational modeling of the effects of the science writing heuristic on student critical thinking in science using machine learning. *Journal of Science Education and Technology*, 1–15

Lee, H. S., Gweon, G. H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09889-7.

Liaw, H., Yu, Y. R., Chou, C. C., & Chiu, M. H. (2020). Relationships between facial expressions, prior knowledge, and multiple representations: a case of conceptual change for kinematics instruction. *Journal of Science Education and Technology*, 1–12

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). Measuring knowledge integration: validation of four-year assessments. *Journal of Research in Science Teaching, 48*(9), 1079–1107. https://doi.org/10.1002/tea.20441.

Maestrales, S. Y., Zhai, X., Touitou, I., Schneider, B., & Krajcik, J. (2021). Using machine learning to evaluate multidimensional assessments of chemistry and physics. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09895-9.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In *Examining pedagogical content knowledge* (pp. 95–132). Springer

McMahon, D., Wright, R., Cihak, D. F., Moore, T. C., & Lamb, R. (2016). Podcasts on mobile devices as a read-aloud testing accommodation in middle school science assessment. *Journal of Science Education and Technology, 25*(2), 263–273. https://doi.org/10.1007/s10956-015-9591-3.

Mercer-Mapstone, L., & Kuchel, L. (2015). Teaching scientists to communicate: evidence-based assessment for undergraduate science education. *International Journal of Science Education, 37*(10), 1613–1638. https://doi.org/10.1080/09500693.2015.1045959.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher, 23*(2), 13–23

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice, 25*(4), 6–20

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement, 53*(3), 265–292

National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. 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. https://link-springer-com.proxy1.cl.msu.edu/content/pdf/10.1007%2Fs10956-011-9300-9.pdf.

Neumann, K., & Waight, N. (2020). The digitalization of science education: Déjà vu all over again? *Journal of Research in Science Teaching, 57*(9), 1519–1528. https://doi.org/10.1002/tea.21668.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

NGSS Lead States. (2018). *Criteria for procuring and evaluating high-quality and aligned summative science assessments*. https://www.nextgenscience.org/sites/default/files/Criteria03202018.pdf.

Nicolaidou, I., Kyza, E. A., Terzian, F., Hadjichambis, A., & Kafouris, D. (2011). A framework for scaffolding students' assessment of the credibility of evidence. *Journal of Research in Science Teaching, 48*(7), 711–744

Osborne, J. (2010). Arguing to learn in science: the role of collaborative, critical. *Science, 1183944*(463), 328

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821–846

Pellegrino, J. W. (2013). Proficiency in science: assessment challenges and opportunities. *Science, 340*(6130), 320–323. https://science.sciencemag.org/content/340/6130/320.long.

Pellegrino, J. W. (2018). Sciences of learning and development: some thoughts from the learning sciences. *Applied Developmental Science*, 1–9

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. ERIC

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the Next Generation Science Standards*. ERIC

Penfield, R. D., & Lee, O. (2010). Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching, 47*(1), 6–24

Rosenberg, J. M., & Krist, C. (2020). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, 1–13

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., & Fortus, D. (2009). Developing a learning progression for scientific modeling: making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632–654

Shavelson, R., Fu, A., Kurpius, A., & Wiley, E. (2013). Evidence- based practice in science education. In *Encyclopedia of Science Education* (pp. 1–4). https://doi.org/10.1007/978-94-007-6165-0_158-1.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: a collaboration between curriculum and assessment developers. *Applied measurement in education, 21*(4), 295–314

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational measurement: issues and practice, 37*(1), 21–34

Sung, S. H., Li, C., Chen, G., Huang, X., Xie, C., Massicotte, J., & Shen, J. (2020). How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology*, 1–17

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2020). Automated scoring of Chinese grades 7–9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology*, 1–14

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14

Wilson, M. (2005). Constructing measures. *An Item Response Modeling Approach*. https://doi.org/10.4324/9781410611697.

Yoo, J., & Kim, J. (2014). Can Online Discussion Participation Predict Group Project Performance? Investigating the Roles of Linguistic Features and Participation Patterns. *International Journal of Artificial Intelligence in Education, 24*(1), 8–32.

Zhai, X. (2019). Applying machine learning in science assessment: Opportunity and challenges. A call for a special issue in *Journal of Science Education and Technology*. https://doi.org/10.13140/RG.2.2.10914.07365.

Zhai, X. (2021). Advancing automatic guidance in virtual science inquiry: from ease of use to personalization. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-020-09917-8.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020a). From substitution to redefinition: a framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658.

Zhai, X., Haudek, K. C., Stuhlsatz, M. A., & Wilson, C. (2020b). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation, 67,* 100916

Zhai, X., Krajcik, J., & Pellegrino, J. (2020c). On the validity of machine learning-based Next Generation Science Assessments: a validity inferential network. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09879-9.

Zhai, X., Li, M., & Guo, Y. (2018). Teachers' use of learning progression-based formative assessment to inform teachers' instructional adjustment: a case study of two physics teachers' instruction. *International Journal of Science Education, 40*(15), 1832–1856

Zhai, X., Shi, L., & Nehm, R. (2020d). A meta-analysis of machine learning-based science assessments: factors impacting machine-human score agreements. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09875-z.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020e). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111–151

Zhang, M. (2015). Google photos Tags Two African-Americans As Gorillas Through Facial Recognition Software. Retrieved on January 3, 2021 from https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=74cae6e1713d.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.