Commentary—Applying Machine Learning in Science Assessment: Opportunity and Challenges

Joseph S. Krajcik¹

Accepted: 15 January 2021 / Published online: 17 February 2021
© The Author(s), under exclusive licence to Springer Nature B.V. part of Springer Nature 2021

Keywords Machine Learning · Knowledge-in-use · Measure student learning

Science education worldwide is at a very exciting and challenging time. Machine learning (ML) holds promise in supporting reforms by helping to improve evaluating large-scale assessments that focus on complex scientific reasoning and use-of-knowledge. The collection of manuscripts in this special issue intersect advances in technology, science education, and assessment to provide insights on how we can make progress using ML to evaluate and provide feedback to knowledge-in-use tasks and further research in designing learning environments to promote knowledge-in-use. For science education researchers and policy researchers who know little about ML, and would like an introduction to ML, I encourage you to read the various manuscripts in this special issue as they provide a clear explanation of what ML is, the various techniques of ML, and the potential and challenges of using ML.

Many countries throughout the world are promoting learning environments and assessments in which learners make use of knowledge rather than just the recall of knowledge (Pellegrino & Hilton, 2012). In the Wang and colleagues' manuscript, *Automated Scoring of Chinese Grades 7–9 Students' Competence in Interpreting and Arguing from Evidence* (this issue), they present results from exploring Chinese middle school science students composing arguments for assessments purposes. This approach contrasts greatly to previous assessment practices in China. PISA (OECD, 2016) for many years has pushed the use of knowledge, but there is only so much that can be done with various forms of selected response items and limited open-end responses. Germany (Kulgemeyer & Schecker, 2014), Finland (Finnish National Board of

Education, 2015), and the USA (NRC, 2012; NAS, 2019) serve as other examples of nations focused on learners developing and measuring competencies. Competencies refer to learning goals or standards expressed in a manner in which learners need to apply their knowledge rather than just recalling knowledge. To further illustrate the change occurring in the Chinese Educational system, in fall of 2019, I attended a conference in China where the Chinese Ministry of Education discussed how to promote creativity and innovation in their K-12 schools (Ministry of Education, P. R. China, 2018). Creativity within the disciplines only occurs when individuals are able to use their knowledge. In the USA, the *Framework for K-12 Science Education* (NRC, 2012) and the Next Generation of Science Standards (NGSS Lead States, 2013) focus on learners using disciplinary core ideas, scientific and engineering practices, and crosscutting concepts to make sense of phenomena or to solve complex problems. This process is referred to as three-dimensional learning by the US science education community.

For years, science education researchers, learning scientists, and cognitive science researchers have tried to figure out how to promote and measure transfer of knowledge or to phrase it in more current turns—individuals being able to use their knowledge to make important personal and societal decisions, make sense of compelling phenomena, solve complex problems, and learn more when needed (Pellegrino & Hilton, 2012). But as a science education community, we have failed this mission for the vast majority of individuals in the world. Yes, some of us, for a variety of reasons, have learned to make use of our knowledge, but for the vast majority of students in K-12 schools and colleges, learning has been a process of memorization and the recall of those ideas on various types of selected response items. The result has been a collective disinterest in many learners, particularly those from disadvantaged backgrounds,

✉ Joseph S. Krajcik
krajcik@msu.edu

¹ CREATE for STEM, Michigan State University,
East Lansing, MI 48824, USA

to pursue science degrees. Many science educators have tried to create innovative learning environments and assessments, but challenges of appropriate curriculum, assessments that are open ended and require students to use their knowledge, sustained professional learning, and resources have not existed for a variety of schools globally.

The current worldwide effort to focus on competencies is an attempt to rectify this situation. A global population that can make informed decisions using knowledge and applying evidence and reasoning is needed. Take for instance, the decision to social distance and wear masks to prevent the spread of Covid-19. Understanding what is a virus, how it spreads, and the risks involved are critical factors in individuals making decisions on how to behave. Should one go to work? Should one go to a bar or out to dinner? Should one visit relatives in a different region than one's hometown? These are choices that each individual needs to make, but the decision needs to be an informed one, based on knowledge and evidence and applying that information, and not just gut reactions. And while the Covid-19 situation is on my mind as the cases soar out of control in many countries throughout world and particularly in the USA, the Covid-19 example is just one example of individuals needing to make informed decisions. Individuals of all ages need to make informed decisions in other areas as well. For instance, making the decision to use public transportation or drive to work. How do various modes of transportation impact air quality and sustainability? In our work lives, individuals will also need to use knowledge and be innovative (National Science Board, 2019). Scientists, engineers, and technicians are needed to design new batteries to

power various electronic devices, and they also need to figure out how to dispose of and recycle batteries for sustainability. How do we go about making sure that we prepare learners to have the useable knowledge for the world in which they live?

The Importance of an Aligned Educational System

In a 2006 report from National Research Council in the USA, *Systems for Science Assessments* (National Research Council, 2006), a systems model was employed to show what impacts student learning (see Fig. 1). Like any system, if you disrupt any component of the system, the system will breakdown or not function as intended. All components in the educational system must work together to promote learning. As the model shows, student learning outcomes will depend upon standards. As discussed above, nations throughout the globe are concerned with focusing on the development of usable knowledge and have developed standards to address this focus. The development of rigorous standards that focus on knowledge-in-use, like the NGSS, might not be perfect, but in many respects are the first step forward to designing an educational system that promotes useable knowledge. But standards are only the first step. If we hope to reach our goal of helping students to use their knowledge, we need to design curriculum, develop professional learning opportunities for teachers, and design assessments that align with knowledge-in-use standards. However, evaluating student responses in assessments that focus on using knowledge required for measuring knowledge-in-use is time consuming, expensive,

Fig. 1 The value of standards (Modified from *Systems for State Science Assessments*, NRC 2006)

and challenging. This major hurdle of evaluating complex and opened ended assessments is addressed in many of the manuscripts in this special issue.

Standards built to promote knowledge in use, such as the Next Generation Science Standards or the new Finish standards, and designed on promoting competencies, provide new assessment and measurement challenges. For instance, the Next Generation Science Standards designed standards, called performance expectations because the focus on performance rather than recall, include all three dimensions of scientific knowledge—disciplinary core ideas, scientific and engineering practices and crosscutting concepts—integrated together. Designing, developing, and scoring assessment tasks that include all three dimensions provide new challenges. Most designers of tasks have traditionally focused on measuring content separate from practice, but assessments that measure knowledge-in-use needed to be performance based, like the NGSS that requires the integrations of all three dimensions of scientific knowledge, providing new challenges for assessment designers. It also raises issues for the design of rubrics and scoring three dimensional items.

The 3-dimensional nature of NGSS assessment tasks raises validity concerns. The scoring of these tasks by ML (Zhai, 2019) raises additional validity concerns. Zhai and colleague in “On the validity of ML-based Next Generation Science Assessments: a validity inferential network,” discuss several validity concerns with evaluating three dimensional learning with machine scoring, including (1) the potential risk of misrepresenting the construct of interest, (2) potential confounders due to more variables involved, (3) nonalignment between interpretation and use of scores and designed learning goals, (4) nonalignment between interpretation and use of scores and actual learning quality, (5) nonalignment between machine scores and rubrics, (6) limited generalizable ability of machine algorithmic models, and (7) limited extrapolating ability of machine algorithmic models. To address these issues, Zhai and colleagues propose a validity inferential network that attends to the cognitive, instructional, and inferential validity of ML-based NGSS assessments.

To accurately measure three-dimensional assessments, complex, open-ended assessment tasks are needed (Harris et al., 2019; Kaldaras et al., 2020), which raises challenges to efficiently and reliably measuring such tasks. When it comes to assessment tasks that state departments of education need to distribute and use, the challenge of evaluating tasks that are three-dimensional can seem daunting. Think of the challenge a state department of education would face to evaluate hundreds of thousands of open-ended three-dimensional tasks. Many state departments of education in the USA have assembled design teams that worked creatively to design assessments that measure knowledge-in-use, but

there is only so much that can be done with selected response items. The State Department of Education in Michigan has designed selected response items to engage learners in designing and developing models that explain phenomena where students select predetermined variables and links to build models. I commend them for this innovative effort, but the cognitive requirement of this drag and drop modeling task is less than what would be required for a student to construct a model without such support.

Although the challenges for states remain large, challenges for classroom teachers also exist. Imagine providing feedback on 100 three-dimensional written explanations, models, or investigation designs even just once a week. The task is daunting even for the most dedicated teacher. This is not only true for K-12 schools but also for large college lectures (see manuscripts in the special issue by Jescovitch and colleagues, and Bertolini and colleagues).

Selected response items are easier and less time-consuming to evaluate, but selective response items do not necessarily provide the cognitive challenge that would be required of students to demonstrate knowledge-in-use. States, classroom teachers, and university instructors face a difficult choice: use various types of selected response items or use constructive response items. Selective response items are typically considered more difficult to write but are easier and less time consuming to score. However, selective response tasks do not measure what we most highly value—challenging cognitive engagement and use-of-knowledge. Constructed response items, on the other hand, are considered easier to construct but more challenging and more time consuming to evaluate, but measure what we most value—engaging learning in more challenging cognitive activities (higher order thinking) where students need to use their knowledge. I do not agree with the argument that constructed response items are easier to develop (see Harris and colleagues, 2019) nor that selected response items cannot measure deeper cognitive levels—they can, to a certain extent, but not to the level of constructed response tasks (and the ones that do are extremely challenging to construct). If we want to make progress to promote learners using their knowledge, then we need to make sure we develop and test items that measure knowledge-in-use (see Harris et al., 2019).

As the model in Fig. 1 shows, removing any one part of the system will not result in the desired effects of student learning. Assessment tasks that measure what we most cherish and need—tasks that promote students using knowledge—are necessary for the system to function to promote learners developing usable knowledge. If the standards are written to promote knowledge-in-use but the assessment tasks measure recall, the system is broken. Teachers and students will respond to what is being measured (the items on the state examine) and not

what is called for. For the system to function, all parts need to be aligned. Professional development and learning environments (curriculum and instruction) that engage students in using their knowledge and providing them with appropriate feedback are also needed. The work by Lee and colleagues in this issue demonstrates one possible learning environment to promote useable knowledge.

The Value of Machine Learning

Using Machine Learning to Evaluate Complex Open-Ended Assessments

Manuscripts in this special issue discuss various methods and the potential for ML to relieve human scoring while at the same time obtaining reliability similar to that of human scorers. The manuscripts by Jescovitch and colleagues and Maestres and colleagues as well as others explored methods that could help improve large-scale ML scoring of open-ended assessment tasks to assess complex use-of-knowledge. However, while ML holds promise in evaluating assessments that focus on students using knowledge, many challenges still remain. Researchers still need to design assessment items that measure what we hope students will learn which can be challenging (see Harris et al., 2019) and evaluation of complex open-end tasks needs to be completed in a timely and efficient manner. The field of ML can help in this area, but it is at its infancy, particular with respect scoring representations and very complex open-ended three-dimensional written tasks. One of the major benefits of this special issue is that it shows that ML can analyze assessment tasks in which students use their knowledge. The set of manuscripts in this special issue illustrates where we are with scoring open-ended assessment tasks and can help the broader science education community understand what ML is (see for instance manuscripts by Jescovitch and colleagues and Lee and colleagues in this issue).

A major advantage of ML is automated scoring of student-generated text in responding to tasks that require learners to provide explanations and arguments to make sense of phenomena, text that describes the design of science investigations (Maestres and colleagues, this volume), and potentially diagrams that learners construct. A valuable finding reported in the special issue is that ML not only works for English language assessment tasks but also for scoring assessment tasks written in Chinese (Wang this volume). Particularly impressive is that the results from scoring the Chinese open-ended tasks, as is true for scoring English language assessment tasks, have similar reliabilities as when human scorers are used to score the tasks. Although the upfront work to train the

machine is substantial, once the machine is trained, thousands of tasks can be easily scored. Considering the growing number of Spanish speaking students in the USA and across the globe, future research to test if ML can evaluate tasks written in Spanish with the same degree of reliability is needed. This will allow researchers to more easily evaluate similar tasks written in English, Spanish, and Chinese.

Developing and using complex constructive response assessments in a classroom environment, whether a K-12 or college context, can provide teachers and instructors as well as students with information to make educational decisions for student learning. However, this feedback needs to occur in a timely manner. The longer it takes to evaluate assessment tasks and provide worthwhile feedback, the less valuable it will be to promote student learning (National Research Council, 2001). In large college instructional contexts, ML can provide valuable, instantaneous feedback (see the manuscripts by Jescovitch and colleagues and Bertolini and colleagues, this issue) to instructors and students on cognitively challenging tasks that go beyond tasks that can be answered by selecting a response. Because of the large class sizes of most introductory college lectures, ML can impact changes in teaching and learning. Jescovitch and colleagues in *Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression* (this special issue) explored the use of ML to automate the scoring of constructive response assessments designed to elicit complex reasoning aligned to a physiology learning progression for undergraduate students. Bertolini and colleagues in “*Testing the Impact of Novel Assessment Sources and Machine Learning Methods on Predictive Outcome Modeling in Undergraduate Biology*” (this issue) investigated various ML techniques and how they could be used to develop models that could provide intervention in large college courses to help eliminate attrition among at-risk students. Because retaining talent from diverse groups in STEM education is so critical (National Science Board, 2019), further research in ML and the creation of models to support at-risk students is a critical area of research.

Use of Machine Learning in Science Education Research

ML holds other potential for improving student learning in science education. ML can also be used to find patterns in research data. The Liaw and colleague manuscript, “*The Relationships between Facial Expressions, Prior Knowledge, and Multiple Representations: a Case of Conceptual Change for Kinematics Instruction*” (this

issue), illustrates the value of ML for tracking facial expression to measure student engagement while performing investigations. This in situ technique can inform us about the emotional state of children. When learners are emotionally engaged, they will more likely invest cognitive energy to work on complex tasks (Csikszentmihalyi, 2008). What is particularly valuable about Liaw and colleagues' research is that second-hand experiences (videos) do not produce the emotional responses as do first-hand experiences. While further research is needed, this finding supports that additional effort and expense to engage learners in direct experiences are critical to promoting engagement. Although much anecdotal evidence supports this claim, the Liaw and colleagues research provides compelling evidence to support this case.

Lee and colleagues in their manuscript, "*Machine Learning-enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments based on Data Drawn from Simulation*" explored how ML automated feedback seamlessly integrated into online curriculum can influence students' performance. In their research, ML was used to score students' use of simulation tasks, and depending on how students responded, different feedback was provided. What is valuable about this work is that it shows the importance of providing students with immediate feedback, which only machine scoring can provide. Additional research and in a variety of different contexts need to further substantiate this promising effort.

Rosenberg and Krist in "*Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations*" investigated how to use unsupervised machine scoring to analyze large qualitative data sets to reveal patterns of the construct under study. As these methods become further refined, they can help researchers identify underlying patterns in large qualitative data, enhancing qualitative research. Their approach simultaneously makes the methodological decisions guiding grounded theory more transparent and reproducible, and it provides more stringent guidelines for qualitatively interpreting meaningful and valid patterns in data using ML methods. However, more research using these ML techniques is needed.

Sung and colleagues (this issue) in "*What is the affordance of machine learning in promptly assessing a student's multimodal representational thinking in an AR-assisted lab?*" investigated ML techniques to code multimodal representational thinking in learners' written representations of lab reports. They compared two ML techniques: one which is commonly used and the other newer deep learning technique. Like the Rosenberg and Kriste, Sung and colleagues used ML to search for themes in a large qualitative data set. Sung and colleagues' claim is that deep learning, a subset of ML approach, is more

accurate in finding patterns than traditional ML methods. Although further research is needed, the deep learning techniques explored in this research holds promise.

The Lamb and colleagues manuscript "*Computational Modeling of the Effects of the Science Writing Heuristic on Student Critical Thinking in Science Using Machine Learning*" illustrated that ML can serve as a significant resource for testing educational interventions to assist in the design and development of future research designs in science-education. The Lamb and colleague manuscript also illustrate that ML can analyze the work of 4th and 5th grade students. This is a critical finding as it shows the ML techniques are applicable across grade bands—elementary through college. Additional research at the elementary, middle, secondary, and tertiary levels is needed.

Concluding Comment

What is so valuable about the studies reported in this special issue on ML in science assessment (Zhai, 2019) is that they show the potential of ML to reliably evaluate complex open-ended assessment tasks and provide almost immediate or just in time feedback to researchers, teachers, instructors, and students on assessment tasks that are valued the most—complex, open-ended assessment tasks that provide information on how learners can use their knowledge. With immediate results teachers and instructors can tailor feedback to differentiate instruction to promote learning. This goes beyond scoring and providing generic feedback to selective response items to students, teachers, and instructors and then providing the most developmentally appropriate assessment task. While this is a valuable step forward and does support differentiation, the potential to provide more meaningful feedback to open-ended tasks exists. Feedback can be tailored to a student's response, promoting the individual to develop deeper levels of understanding. While there is much work to be done to produce reports that teachers, instructors, and students can easily interpretate, imagine the tailored scaffolding that could be developed to support learners. What is incredible is that the work presented in this special issue does not focus on interpreting selective response items but in analyzing student written responses. And as pointed out in the Sung and colleague manuscript, these techniques may also help to identify less engaged learners. Although it will take time before these ML techniques become available to teachers, instructors, and students at scale, the potential to promote knowledge-in-use is staggering. Solving the challenge of how to efficiently evaluate and provide quality feedback is one hurdle that educators throughout the globe will need to solve to produce educational systems that will develop individuals who can use their knowledge to solve complex

problems, make decisions and learn more when needed. The assessment component of educational systems is a critical component that ML, as the manuscripts in this special issue demonstrate, can help to solve.

ML also holds promise to support college instructors, teachers, and researchers in modifying materials and instruction for future uses. Using ML researchers could more rapidly identify themes in larger, qualitative data sets which can provide insights in how to modify learning environments. As several of the manuscripts have discussed, patterns or themes to classify student responses can be immediately available allowing researchers and designers to more quickly make changes to design of curriculum and software. As a researcher who engages in the design of learning environments, receiving such immediate information, coupled with observing classrooms, would provide for much more agile development efforts. Overall, advances in ML are allowing science education researchers to analyze students' complex performances captured in open-ended text responses that transform the teaching and learning of K-16 science education. Such research can help to improve assessment, instruction, and curriculum materials to promote student learning (see Fig. 1).

References

- Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. New York: Harper Perennial.
- Finnish National Board of Education (FNBE). (2015). National core curriculum for general upper secondary schools 2015. *Helsinki, Finland: Finnish National Board of Education (FNBE)*. Retrieved from. http://www.oph.fi/saadokset_ja_ohjeet/opetussuunnitelmien_ja_tutkintojen_perusteet/lukiokoulutus/lops2016/103/0/lukion_opetussuunnitelman_perusteet_2015
- Harris, C. J., Krajcik, J., Pellegrino, J., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2), 53–67.
- Kaldaras, L., Akaeze, H., & Krajcik, J. (2020). Developing and validating Next Generation Science Standards-aligned learning progression to track three-dimensional learning of electrical interactions in high school physical science. *Journal of Research in Science Teaching*, 1–30. <https://doi.org/10.1002/tea.21672>
- Kulgemeyer, C., & Schecker, H. (2014). Research on educational standards in German science education—towards a model of student competences EURASIA. *Journal of Mathematics, Science & Technology Education*, 10(4), 257–269.
- Ministry of Education, P. R. China. (2018). *Curriculum plan for senior high school* [普通高中课程方案]. Beijing: People's Education Press.
- National Research Council. (2001). *Knowing what students know: the science and design of educational assessment*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/10019>.
- National Research Council. (2006). *Systems for State Science Assessment*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/11312>.
- National Research Council (NRC). (2012). *A framework for K–12 science education: practices, crosscutting concepts, and core ideas*. National Academies Press.
- National Academies of Sciences, Engineering, and Medicine. (2019). *Science and engineering for grades 6–12: investigation and design at the center*. Washington, DC: The National Academies Press.
- National Science Board. (2019). *The skilled technical workforce: crafting America's science and engineering enterprise*. National Science Board.
- NGSS Lead State Partners. (2013). *Next Generation Science Standards: for states, by states*. The National Academies Press.
- OECD. (2016). *PISA 2015 Assessment and analytical framework: science, reading, mathematic and financial literacy*. Paris: OECD Publishing.
- Pellegrino, J. W., & Hilton, M. L. (2012). *Committee on defining deeper learning and 21st century skills*. Washington DC: National Academies Press.
- Zhai, X. (2019). Applying machine learning in science assessment: opportunity and challenge. *Journal of Science Education and Technology*. <https://doi.org/10.13140/RG.2.2.10914.07365>.

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