Kaldaras *et al.*
International Journal of STEM Education (2024) 11:57
https://doi.org/10.1186/s40594-024-00516-0

International Journal of STEM Education

COMMENTARY

Open Access

Employing automatic analysis tools aligned to learning progressions to assess knowledge application and support learning in STEM

Leonora Kaldaras[1*], Kevin Haudek[2] and Joseph Krajcik[2]

Abstract

We discuss transforming STEM education using three aspects: learning progressions (LPs), constructed response performance assessments, and artificial intelligence (AI). Using LPs to inform instruction, curriculum, and assessment design helps foster students' ability to apply content and practices to explain phenomena, which reflects deeper science understanding. To measure the progress along these LPs, performance assessments combining elements of disciplinary ideas, crosscutting concepts and practices are needed. However, these tasks are time-consuming and expensive to score and provide feedback for. Artificial intelligence (AI) allows to validate the LPs and evaluate performance assessments for many students quickly and efficiently. The evaluation provides a report describing student progress along LP and the supports needed to attain a higher LP level. We suggest using unsupervised, semi-supervised ML and generative AI (GAI) at early LP validation stages to identify relevant proficiency patterns and start building an LP. We further suggest employing supervised ML and GAI for developing targeted LP-aligned performance assessment for more accurate performance diagnosis at advanced LP validation stages. Finally, we discuss employing AI for designing automatic feedback systems for providing personalized feedback to students and helping teachers implement LP-based learning. We discuss the challenges of realizing these tasks and propose future research avenues.

Keywords Learning progressions, Artificial intelligence, Machine learning, Performance assessments, Knowledge application, STEM

Introduction

Artificial intelligence (AI) holds tremendous potential to transform all fields of human activity, including education, because the AI models can be trained to automate or streamline various processes and potentially can collaborate with or advance human activity. For example, one of the most popular AI tools, ChatGPT, is trained on large amounts of conversational data related to education and, therefore, is capable of considering context and tailoring its responses to the specific needs of the user such as personalizing learning experiences to the needs of individual students (Samala, Zhai, Aoki, Bojic, & Zikic, 2024). However, the AI revolution comes during a period of significant changes in the field of global education itself. Specifically, recent educational reform efforts worldwide emphasize the need to foster knowledge application to support learning in science, technology, engineering, and mathematics (STEM) at the K-12 and the undergraduate level (National Research Council [NRC], 2012; PISA, 2025).

The move towards supporting knowledge application calls for significant changes at all stages of the learning process including the implementation of new learning systems that help foster knowledge application grounded in developmental models of learners. Effective

*Correspondence:
Leonora Kaldaras
Leonora.Kaldaras@ttu.edu
[1] Department of Curriculum and Instruction, Texas Tech University, Lubbock, USA
[2] CREATE for STEM Institute, Michigan State University, East Lansing, USA

development and implementation of such learning systems depend upon the ability to adjust the learning process to meet the needs of academically, culturally, and linguistically diverse learners. Learning progressions (LPs) represent cognitive models that describe how the understanding (competency) of big ideas develops over a broad, defined period (Duschl et al., 2007). Building education around validated LPs can help to meaningfully adjust the learning process to the needs of individual diverse learners and, therefore, contribute to creating more equitable and effective learning systems (Kaldaras & Krajcik, 2024). However, these positive outcomes are contingent upon our capability to streamline every stage of the process, including the development and validation of the relevant LPs aligned to specific standards, carrying out timely and meaningful evaluation of student progress on LP-aligned assessments, providing feedback to individual students and teachers, and supporting meaningful use of this feedback to improve learning outcomes (see Fig. 1). The scope of the LP targeted in this study focuses on standards that describe knowledge application skills reflected primarily in students' ability to integrate relevant disciplinary knowledge and scientific practices when explaining real-life phenomena and solving complex problems. The Next Generation Science Standards (NGSS) represent a well-defined example of such standards at the K-12 level (NGSS Lead States, 2013). However, the approaches to leveraging AI for streamlining the process of LP-based learning are not unique to NGSS and are applicable across various educational settings where the end goal of instruction is to foster useable knowledge of science among learners reflected in applying their understanding to real-life scenarios by engaging in authentic scientific practices and using other relevant skills.

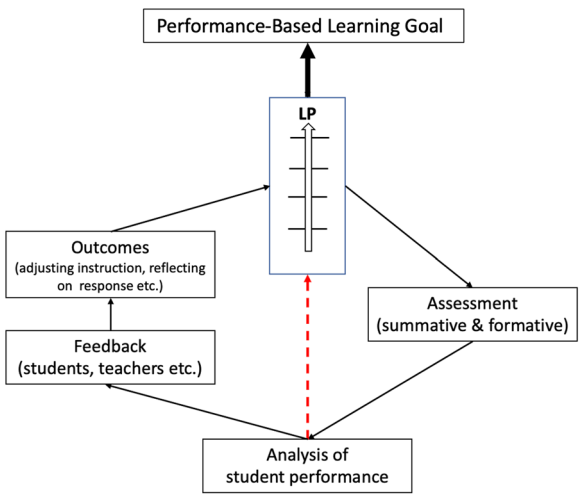

Fig. 1 Learning Progression-based vision of the education focused on adjusting the learning process to the needs of individual learners via timely and meaningful LP-aligned feedback

AI holds the potential to streamline and enhance each stage of the process discussed above and shown in Fig. 1. This paper aims to discuss potential challenges and opportunities associated with using AI at each of these steps and outline further research avenues aimed at leveraging AI for implementing LP-based vision of education into practice. We will consider commonly used AI approaches in education, including supervised, semi-supervised, and unsupervised machine learning (ML) and generative AI (GAI). Supervised ML uses previously labeled data (e.g., previously scored student responses) to predict outcomes (e.g., scores on a new set of student responses). Unsupervised ML uses previously unlabeled data to learn or find patterns without explicit instructions or labels. Semi-supervised ML refers to using partially labeled data sets to predict outcomes. Finally, GAI is a combination of supervised and unsupervised ML that can identify patterns in existing data sets and generate new content with similar characteristics. For example, GPT refers to a family of GAI pre-trained models (e.g., GPT released in 2018, an updated model called GPT-2 released in 2019 etc.). ChatGPT refers to a chat-bot powered by GPT model that provides a simple-to-use interface, making the GPT technology accessible to an average user.

Importance of using learning progressions to foster knowledge application

Recent educational reform efforts worldwide emphasize the need to foster knowledge application to support learning in science, technology, engineering, and mathematics (STEM) at the K-12 and undergraduate levels. These efforts include PISA, which has emphasized knowledge application in their assessment (PISA, 2025). Further, Germany (Kulgemeyer & Schecker, 2014) and Finland (Finnish National Board of Education, 2015) have developed national standards to support learners in developing and measuring competencies. Competencies in this context refer to standards expressed as learning goals that require learners to apply their knowledge rather than reciting back memorized information. A similar push towards measuring competencies is occurring in the Chinese educational system (Ministry of Education & P.R. China, 2020; Yao & Guo, 2018). In the US, similar efforts have resulted in the publication of the Framework for K-12 Science Education (*the Framework*) and the NGSS, which emphasize the importance of fostering knowledge growth coherently over time so learners can apply what they learn (NGSS Lead States, 2013; NRC, 2012). The National Assessment Governing Board

(NAGB) has also released an updated science framework for the 2028 Nation's report card that recommends supporting learners in developing the ability to integrate disciplinary knowledge and scientific practices to foster understanding (National Assessment Governing Board, 2023).

The Framework for K-12 Science Education (*the Framework*) (NRC, 2012) builds on years of research on how students learn (Bransford, et al., 2000; Duschl et al., 2007; Pellegrino et al., 2001) and emphasizes the importance of aligning curriculum, instruction, and assessment along empirically validated learning pathways grounded in relevant cognition theories (NRC, 2006). These learning pathways, called learning progressions (LPs), provide cognition models that describe how the understanding (competency) of big ideas develops over a broad, defined period (Duschl et al., 2007). Building education around validated LPs ensures a valid interpretation of student progress with respect to relevant cognition theories as reflected in performance on LP-aligned assessments (Brown & Wilson, 2011; Mislevy, 1996). If LPs are not used as a guide, it becomes challenging to interpret student progress, which limits the validity of the associated assessment results (Mislevy, 1996). Additionally, a lack of information on how proficiency develops makes it harder for teachers to support their students in developing competencies in a topic.

Learning progressions describe learning as it develops under carefully scaffolded curriculum and instruction (Duncan & Hmelo-Silver, 2009; Kaldaras & Krajcik, 2024; Krajcik & Namsoo, 2023). Thus, LPs are not developmentally inevitable because they do not describe learning as it naturally occurs but rather how learning develops under supportive educational conditions. Additionally, the LPs, such as those described in *the Framework*, reflect scientific reasoning focused on student progression towards a deeper understanding of a topic, as reflected in the ability to apply knowledge when explaining phenomena and solving real-life problems (NRC, 2012). Therefore, LPs represent complex cognitive frameworks that combine elements of content (disciplinary core ideas and crosscutting concepts, NRC 2012) and practice and focus on describing complex understanding necessary to make sense of complex phenomena and solve challenging problems.

Specific challenges are associated with implementing an LP-based education vision focused on developing competencies. Specifically, LPs are hypothetical in nature and require validity evidence to be used in practice. Typically, validating an LP requires developing a hypothetical LP for a construct, designing assessments capable of probing LP levels, and collecting and analyzing response data to ensure that theoretically suggested LP levels are supported in practice. However, it is challenging to measure student progress along the LP describing complex constructs using only multiple-choice (MC) or closed-form items (Krajcik, 2021). This is because measuring proficiency on complex constructs described by a LP requires aligned assessments that enable students to apply their content understanding when using scientific practices at different levels of sophistication. This approach measures proficiency along an LP as reflected in level-specific competencies demonstrated by students. This proficiency is complex and measuring it requires students to respond to complex scenarios to demonstrate their ability to apply their understanding (Krajcik, 2021; National Research Council [NRC], 2014).

The assessments capable of capturing proficiency associated with knowledge application are called performance assessments. In contrast to content-focused assessments, performance assessments require students to use scientific practices while applying relevant scientific ideas to explain phenomena and solve novel and complex problems. The examples of practices include developing explanations using evidence and scientific reasoning, constructing models that provide mechanisms for phenomena, and making arguments supported by evidence (NRC, 2014). Measuring proficiency focused on knowledge application can be achieved using different assessment formats, including MC and constructed response (CR). For example, the facet and construct-centered approaches to MC item design are informed by a model relating student selection of "distractors" on items as reflecting a particular form of understanding (Wind et al., 2019). However, such items do not provide teachers with information regarding students' reasoning, a critical aspect of knowledge application reflected in the ability to go beyond simply recalling information. Therefore, measuring proficiency focused on knowledge application could be done using MC, but most MC items do not do so. CR assessments, like constructing explanations with evidence and reasoning and developing models, provide rich information and require learners to use scientific reasoning. However, since students produce the responses, they have been time-consuming and expensive to evaluate and provide feedback for (Krajcik, 2021). Relatedly, implementing a LP-based vision of education in practice requires that students have learning opportunities to meaningfully engage with the content under study, which, in turn, requires that teachers and students receive timely LP-aligned feedback about student performance. Obtaining timely feedback in practice is time-consuming due to a similar reason—the requirement to quickly and efficiently evaluate student performance on various assessments with respect to LP levels. This task is even more challenging when aiming to achieve equitable

learning outcomes because we need to ensure that the LP and the LP-aligned feedback accounts for the diversity of ways of thinking about a scientific idea at various proficiency levels. Meaningful evaluation of the diversity of student thinking at all LP levels and timely and appropriate feedback that incorporates this diversity to learners and teachers can help create valuable learning opportunities and promote improved learning outcomes among diverse student populations.

Artificial intelligence (AI), such as natural language processing (NLP) and machine learning (ML), has been implemented in a variety of educational contexts for scoring CR performance assessments (Litman, 2016). Prior to the release of GAI technology, such as ChatGPT, one primary use of AI in education was as an assessment evaluation tool, for example, to speed up or scale the scoring of CR items (Chiu et al., 2022; Wilson et al., 2023). With the emergence of GAI, new possibilities for using AI to assist in the design and implementation of LP-based learning environments emerge. We argue that employing AI approaches represents a promising way for efficient LP validation and implementation of LP-based vision into practice more broadly. Validated LPs coupled with the capabilities of AI, in turn, can inform curriculum, instruction, and assessment development. Currently, no complete examples exist that illustrate the use of AI to validate and implement LPs in practice. We further describe the challenges and promises of using AI (both supervised and unsupervised ML as well as NLP and GAI such as GPT) for LP validation and use in classrooms to facilitate a shift towards an education system focused on fostering knowledge application. We will discuss potential research avenues in this direction and describe examples of current projects that aim to achieve this vision.

Challenges of designing and validating LP-based learning systems that support the development of knowledge application ability

Using validated LPs as roadmaps for building learning systems that support knowledge application represents a promising way of implementing the new vision of science education (Brown & Wilson, 2011; NRC, 2014, 2012). Figure 2 is a modification of Fig. 1 and shows the different stages of implementing LP-based vision into practice, starting with the development and validation of LPs and associated assessments, (1) followed by evaluation of student progress of the assessments with respect to the LP levels (2), providing timely and meaningful LP-aligned feedback to students and teachers (3), and fostering meaningful outcomes based on the feedback (4). These stages reflect general practical steps of LP-based education and represent critical time and resource-consuming stages of implementing the LP-based vision of education into practice. We propose that each of these stages can be streamlined with AI, therefore, making the LP-based

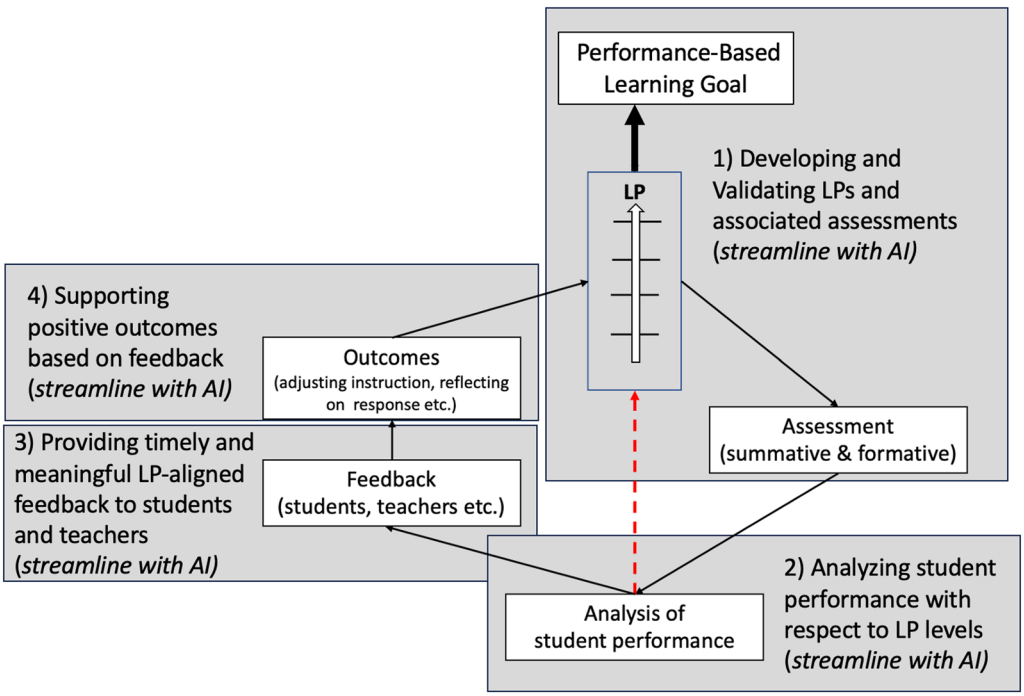

Fig. 2 Stages of implementing LP-based vision of education into practice. We propose that each of these stages can be streamlined with AI

vision of education closer to reality. In this section, we discuss common challenges associated with each of these stages. In the following section, we discuss the challenges and promises of streamlining these steps and addressing the noted issues using AI.

Stage 1: developing and validating LPs and associated assessments

The degree to which an LP can be used to support curriculum, instruction, and assessment depends on the available validity evidence for the LP. The overarching approach to LP validation is grounded in the vision outlined in the Standards of Psychological and Educational Testing (Eignor, 2013; Messick, 1980), which refers to validity as appropriateness, meaningfulness, and usefulness of specific inferences from test scores. In the context of LP validation, test scores refer to student performance on an assessment designed to probe various LP levels. These standards further say that test validation is a process of accumulating evidence to support the inferences. The most common type of validity evidence for LP pertains to construct validity, reflected in evaluating the degree to which student responses on the assessments designed to probe LP levels fall into the hypothetical LP levels. Accumulating such evidence is an iterative process. At each stage, evidence is collected to inform further refinement of the LP levels, revision of the associated assessments, and inform curriculum and instruction as needed.

LP validation studies focusing on construct validity can generally be divided into three stages: (1) developing a hypothetical LP grounded in a review of relevant literature (no assessment instrument or associated curriculum) (e.g., Duncan et al., 2009; Smith et al., 2006); (2) using an existing hypothetical LP and corresponding assessment tasks to investigate how well existing curriculum supports student learning along the progression (includes associated assessment instrument but does not have LP-aligned curriculum) (e.g., Herrmann-Abell & DeBoer, 2018; Mohan et al., 2009); (3) using a previously developed hypothetical LP with LP-aligned assessment tasks and curriculum to investigate how well the curriculum supports student learning along the LP levels (e.g., Lehrer & Schauble, 2000; Songer et al., 2009).

An added challenge for the validation of LPs that describe knowledge application is the complexity of the underlying constructs. Specifically, constructs describing knowledge application ability combine elements of content and practice. Validating LPs for such complex constructs calls for obtaining evidence of students' abilities to integrate relevant disciplinary knowledge and scientific practices when explaining phenomena and solving novel problems (Kaldaras et al., 2021a, 2021b, 2023; NRC, 2012). Approaches to developing and obtaining construct validity for LPs describing complex constructs (Gunckel et al., 2022; Kaldaras et al., 2021a, 2021b; Scott et al., 2022) and for designing performance assessments measuring complex constructs (Harris et al., 2019; Mislevy & Haertel, 2006; San Pedro et al., 2014) have been detailed elsewhere. In previous validation studies, CR performance assessments and interviews (Gunckel et al., 2022; Kaldaras et al., 2021a, 2021b, 2023; Scott et al., 2022) were used to gain insight into student ability to integrate elements of content and practice and demonstrate knowledge application. One of the challenges for construct validation of these LPs is developing valid assessments capable of measuring complex constructs along the LP levels (Kaldaras & Krajcik, 2024). For example, recently, several studies focused on developing and validating LPs aligned to the Next Generation Science Standards (NGSS) (Kaldaras, 2020; Kaldaras et al., 2021a, 2023). For example, Kaldaras et al. (2021a) described the development and validation of NGSS-aligned LP for electrical interactions. They demonstrated how the validated LP can be used to accurately place individual students on an LP level with a high degree of precision. A similar study was conducted for the construct of chemical bonding (Kaldaras et al., 2023).

The challenge of developing and validating knowledge-in-use LPs lies in the need to carefully describe increasingly sophisticated ways of integrating relevant disciplinary knowledge with scientific practices and other important skills when applicable. For example, in the context of NGSS, validating LPs calls for carefully specifying increasingly sophisticated ways of integrating the three dimensions of scientific knowledge—disciplinary core ideas (DCIs), scientific and engineering practices (SEPs), and crosscutting concepts (CCCs)—at varying levels of sophistication, and design assessment instruments capable of probing the levels of such an LP with a high degree of accuracy and precision. As described above, assessments that probe levels of such an LP would require students to engage in relevant scientific practices and demonstrate the ability to apply science content (DCIs and CCCs) to explain phenomena. This often calls for CR assessments. In this context, CR assessments refer to a range of response types that reflect student ability to integrate relevant disciplinary knowledge (DCIs and CCCs) and practices, which is indicative of complex reasoning. For example, if a given assessment measures students' ability to develop scientific explanations of relevant phenomena using underlying big ideas in science, a CR response would reflect a short text-based response demonstrating students using relevant disciplinary knowledge to develop a causal scientific explanation of the phenomenon in question. CR scientific explanations represent one of the most common types of responses

previously studied to be evaluated with AI (e.g., 2022b; Kaldaras et al., 2022a; Wilson et al., 2024). Yet another fundamentally different CR response type reflects scientific models that call for learners to develop representations (e.g., drawings) that explain phenomena as opposed to text-based CRs that can be used to evaluate scientific practices of developing explanations and arguing from evidence. Recent efforts have begun to investigate applying AI to evaluate modeling CRs (e.g., Kaldaras et al., 2024a, 2024b, 2024c; Sagherian et al., 2022).

Considering the time and resources it takes to develop, validate, and evaluate student performance on these assessments with respect to LP levels, the LP validation becomes an incredibly resource and time-consuming process. The same challenges apply to developing and validating LPs describing complex constructs that combine elements of disciplinary knowledge and practice beyond NGSS-related topics (e.g., Kaldaras & Wieman, 2023a, 2023b). However, effectively supporting the development of knowledge application ability becomes challenging if we lack validated LPs that can guide how students develop proficiency in complex constructs. We posit that AI can help streamline LP development and validation, including quick and efficient evaluation of information provided by these LP-aligned assessments and identifying response patterns that can serve as evidence for LPs.

Stage 2: analyzing student performance with respect to LP levels

Evaluating information from LP-aligned performance assessments gives rise to several issues stemming from student understanding being context-dependent and complex (Alonso & Elby, 2019; Hammer & Sikorsky, 2015; Sikorsky, 2019). Specifically, understanding does not necessarily progress linearly and cannot always be described by a singular LP path. Each learner brings their own unique experiences and knowledge, and therefore, all start at different levels of understanding, which makes it challenging to define a lower anchor of an LP. Similarly, a clear-cut definition of an upper anchor remains a challenge (Sikorsky, 2019). Further, as students participate in learning, their understanding still depends on specific learning and instructional strategies they experience, as well as prior experiences. This gives rise to what is termed in the literature as the "messy middle" of an LP (Alonso & Elby, 2014; Hammer & Sikorsky, 2015). The "messy middle" reflects various ways students can learn and apply a given construct, and that cannot necessarily be arranged neatly in increasing sophistication (Alonso & Steedle, 2009; Shavelson & Kurpius, 2012; Sikorsky, 2019).

The issue of the "messy middle" is likely even more significant for describing complex constructs since such constructs combine elements of disciplinary knowledge, as well as various aspects of scientific practices, CCCs and other skills (Gorin & Mislevy, 2013; Kaldaras, et al., 2021b). For example, applying the disciplinary ideas related to electrical interactions when developing scientific models of phenomena could look very different depending on the specific context (e.g., ionic forces in chemistry versus point charge in the electric field in physics). Even within the same assessment scenario, students can demonstrate their understanding in many ways (Alonso & Steedle, 2009; Kaldaras et al., 2023). Making sense of these differences in student reasoning with the same disciplinary ideas and practices and how they relate to LP levels (if at all) is critical for obtaining construct validity evidence for the LP. Evaluating a sufficiently large number of diverse student responses to LP-aligned assessments is critical for building a compelling validity argument that accounts for the diversity of student thinking at all levels of sophistication. Achieving this in practice is time and resource consuming. Leveraging AI technology such as unsupervised ML can help identify various specific patterns in student thinking, evaluate those patterns with respect to LP levels, and offer strategies for revising LP level descriptions during the iterative validation cycle to account for the diversity of thinking patterns. GAI can also identify relevant patterns and provide an initial description of the categories under which these patterns fall. A recent study by Kaldaras and colleagues described leveraging GAI to identify patterns in student responses to develop rubrics for LP-aligned assessments (Kaldaras, et al., 2024). However, similar approaches can be used at the LP validation stage, as discussed here as well.

Streamlining this process by leveraging AI can potentially offer a meaningful way of tackling the issue of the "messy middle" by providing a quick way to identify patterns that do not align well with LP and streamline the validation process by either informing LP revision or finding alternative ways of dealing with atypical patterns if they cannot be meaningfully incorporated into the LP. If such patterns are identified, this does not diminish the value of the LP but instead provides additional insights on how student understanding develops and evolves based on prior knowledge and experiences, which will help adapt the learning process to the needs of individual diverse learners and create more equitable learning environments as a result.

Stages 3 and 4: providing LP-aligned feedback and supporting positive learning outcomes based on the feedback

The discussion so far has been focused on the construct validity of LPs. However, to successfully use LPs

in practice, validity evidence should be obtained on how well an LP supports its intended use. This validity is termed consequential validity (Alonso & Elby, 2019). Demonstrating consequential validity requires formulating a theory of action aimed at teachers that outlines the ways an LP should be used, accompanied by evidence, demonstrating that the LP promotes learning when used in those ways (Alonso & Elby, 2019). We believe that AI approaches can demonstrate consequential validity and support using LPs in practice. We further discuss using AI approaches for tackling issues pertaining to construct and consequential validity and describe potential challenges associated with using AI for these purposes.

Implementing LP-based vision using AI approaches: promises and challenges

We now discuss leveraging AI to tackle the challenges of implementing an LP-based vision of science education in practice, as discussed above. We will specifically focus on discussing promises and challenges of using AI at each stage shown in Fig. 2.

Stage 1: leveraging AI for developing and validating LPs and associated assessments

Promises

As mentioned above, at the initial stage of LP validation (construct validation), research focuses on developing a hypothetical LP typically grounded in relevant literature, conversation with educational and disciplinary experts, and an overview of the available research. There are very few hypothetical LPs that describe complex constructs describing knowledge application available in the literature. A central challenge in formulating these hypothetical LPs is recognizing distinctly different ways that students can reason with a construct, accounting for the diversity of student thinking at each level of sophistication and determining the order of sophistication for those various reasoning patterns if such an order exists. Traditionally, this is achieved by researchers spending considerable time and resources studying available information to identify such patterns and orderings. With the emergence of AI technology, this process can be streamlined using a computational grounded theory approach (Nelson, 2020). Such a framework utilizes a strength of AI techniques to identify patterns in large datasets. These patterns are then subsequently refined and confirmed by experts. Specifically, some studies have used AI approaches such as ML or learning analytics, to identify patterns in the available data (Berland et al., 2013; Gobert et al., 2013). GAI could also offer important insights into the types of reasoning and skills and their order for the initial development of an LP. Researchers can evaluate the emergent patterns and their importance in terms of pedagogy and/or student learning. Moreover, in cases when a sufficiently diverse data training set is not available to capture a wide range of diverse student responses reflecting varying levels of sophistication in a given construct, GAI can be leveraged to generate responses reflecting additional diversity. For example, a recent study by Martin and Graulich (2024) shows that combining authentic human-generated and GAI chatbot-generated responses yields the highest human–machine agreement across validation conditions in capturing student causal reasoning about the mechanism of chemical reactions.

Further, a more advanced stage of LP development for complex constructs relates to developing performance assessments probing levels of the hypothetical LPs. When there are no available hypothetical LPs, one way to start developing an LP could be to use existing AAAS maps (Mccomas, 2014), NGSS strand maps (NGSS Lead States, 2013) or the NSTA Atlas of the Three Dimensions (Willard, 2020) that outline what students should be able to do to demonstrate proficiency in a construct. Such strand maps are often more readily available from prior work or more easily developed through analyzing a content domain. Performance assessments aligned to these maps can further be developed to measure student proficiency in a domain with subsequent use of AI approaches for identifying patterns in student reasoning. In cases where specific relevant performance standards are not readily available, one could leverage GAI to identify patterns across various standards potentially relevant to a given topic and use these patterns to develop new standards for the topic of interest. For example, a recent study developed a GAI-based approach that guides alignment among the various standards by reducing the number of potential pairs subject matter experts need to consider when aligning the standards to only those that should be considered due to high semantic overlap (Butterfuss & Doran, 2024).

Further, pattern recognition at the initial stages of LP development can be achieved by using unsupervised or semi-supervised machine learning (ML) following a computational grounded theory approach (Nelson, 2020). Unsupervised ML attempts to find patterns and associations in data and does not require previously labeled data (Kotsiantis, 2007), which is often unavailable at the initial stages of LP development. Semi-supervised ML attempts to find patterns in partially coded data (Kotsiantis, 2007). These approaches can find large-scale patterns in the available data, potentially shortening the development time of LPs by helping researchers identify common ways students think about topics and produce frameworks or coding rubrics based on patterns in the data. Recent work has employed unsupervised ML along with qualitative analysis and pattern finding during the

development of a construct map for students' model-based explanations (Rosenberg & Krist, 2020). Other studies have used common unsupervised ML techniques, like data clustering or NLP, to provide some validity for qualitative coding (Sherin, 2013) or as part of a mixed methods approach to develop better ML predictive models (Sripathi et al., 2023; Tschisgale et al., 2023; Wiley et al., 2017) or to identify patterns or themes within text (Chang et al., 2021). Additionally, recent studies on initial LP validation for complex constructs focused on identifying relevant response patterns either based on written student responses or interviews (Gunckel et al., 2022; Scott et al., 2022). In these studies, such analysis is done by researchers and requires a significant amount of time and resources. Trained GAI systems can also recognize relevant patterns in student response data and help develop the progression of student understanding of a given construct. For example, similar studies on leveraging GAI to develop a progression have been conducted in medicine to determine progression for various types of cancer treatment (Zaballa, Pérez, Inhiesto, Ayesta, & Lozano, 2023). Further, unsupervised ML and NLP could potentially assist this process by providing quick pattern identification, which researchers analyze for their relevance for LP validation. In a recent study using such an approach and leveraging recent approaches to deep learning, Martin and colleagues (2023) identified dozens of response clusters in college student writing in chemistry. These identified patterns were interpretable and linked to frameworks related to reasoning and granularity to explain the complexity of the response. This represents an application of AI during the process of LP development, where AI outputs are used to assist the development instead of only being used as an endpoint of assessment (Kubsch, et al., 2022).

Since the initial early stages of LP development do not have associated aligned assessment and curriculum materials that afford greater coherence in student thinking, student reasoning patterns can be highly diverse, making it challenging to identify the relevant patterns, including the "messy middle." Research could use NLP, unsupervised ML, and GAI as a helpful starting point for identifying reasoning patterns from a more significant number of responses to various performance assessments. While unsupervised ML might not provide "the solution" to the "messy middle" and the diversity of student thinking, it may offer approaches to help better understand the "messy middle." For example, results from this approach may identify multiple, distinct ways of reasoning used by sub-groups of students who would be characterized within the "messy middle", thus providing a different grain size to begin to unpack student reasoning and learning. Similarly, if identified patterns over the entire LP do not seem to support an increasingly sophisticated pathway, this suggests that a construct might not be described by a linear LP (Alonso & Elby, 2019). The identified patterns, however, could offer insights into other possible, non-linear ways of characterizing proficiency in a construct, thereby making the development of a cognition model for that construct more plausible. For example, no LP-based cognition model currently exists for the complex construct of problem-solving (Adams & Wieman, 2015; Price et al., 2022), and it is unlikely that a linear LP framework cannot meaningfully describe this construct. However, sufficient literature exists on different ways students engage in problem-solving tasks across various disciplines (Burkholder et al., 2020, 2021). AI can analyze this literature to identify distinct patterns of engaging in problem-solving practice in various contexts. These findings can further be leveraged to guide the design of learning environments for supporting students in developing problem-solving abilities. Alternatively, these findings could be leveraged to build LPs for various content ideas focused on describing problem-solving proficiency in applying specific content ideas, which could guide learning.

In the context of LP validation for complex constructs, AI could help obtain validity evidence that would distinguish between different but closely related finer-grained LPs. Using the example from above, validating an LP describing students' ability to combine disciplinary knowledge related to electrical interactions with different scientific practices such as modeling, constructing explanations, or arguing from evidence could lead to three different LPs. All those LPs would be related and some combinations might be indistinguishable in proficiency levels. AI approaches offer a quick way to score the associated performance assessments. Once the scores are available, they can easily be used in subsequent analysis to deduce how related the LPs are and whether they are specific enough to assign levels to individual students with a high degree of accuracy. This information is desirable at later stages of LP validation to ensure that an LP can be easily used to inform topic-specific instruction, curriculum, and assessment strategies.

Challenges

Despite promising advances, using unsupervised ML or GAI for initial LP development and validation poses particular challenges. Specifically, in unsupervised or semi-supervised ML, the machine is set to find patterns or derive the logic from a pool of uncoded or partially coded data (Gobert et al., 2015). These present validity challenges since there are no or few human-assigned codes to compare. Instead, the researchers must qualitatively analyze the groupings and predictions to determine if the

predicted patterns align with constructs of interest (Martin et al., 2023). Further, even when unsupervised ML predictions are accurate or relevant, it is not always obvious what logic or features in the algorithms contribute to the machine predictions. That is, we still are unsure that those predictions are due to applying logic or identifying critical features in the same way as a human scorer (Yang et al., 2002). This presents a serious issue for the validity of the resulting scores or classifications and should be considered if using semi-supervised or unsupervised ML at the early stages of LP development to ensure construct validity. Similarly, when GAI, it is important to train the algorithm to use relevant criteria for a construct when identifying patterns (Kaldaras, et al., 2024). One could think of this process as training an apprentice to ensure that they use commonly accepted criteria to carry out the task at hand. The training process should be transparent and verifiable such that we can put trust into the resulting identified patterns and their relevance for the LP of interest.

Stage 2: analyzing student performance with respect to LP levels

Promises

This stage involves researchers using a set of assessment items that probe various LP levels, well-defined scoring rubrics aligned to the LP levels and possibly more information on context-dependency and diversity of student reasoning. In the context of LPs and assessments measuring knowledge application ability, CR assessments are often used. When scoring LP-aligned CR performance assessments, student responses can be automatically scored using NLP, supervised ML, or GAI. The supervised ML approach requires students' responses to assessment items with codes or scores, aligned to various science ideas in the LP, assigned by expert raters. The labeled data is used to train the ML program to recognize these ideas and subsequently serve as an automatic scoring method that does not require a human scorer after the program has achieved a sufficient level of human–machine agreement. This approach is referred to as *supervised machine learning* because it requires a training set to be supplied to the program initially, and it has been widely used to predict student CR scores and forms the basis of many automatic scoring approaches. There are both technological, design-based, and validity challenges associated with using AI to evaluate student performance with respect to LP. All three will be further discussed below.

Technological challenges

In the case of supervised ML, the technological challenge is that supervised ML model development is initially time-consuming because it needs a large sample of responses to be labeled for training (typically, about 1000 responses per item). Since responses are item-specific, one ML model usually fits only one item, and items and, therefore, models are not easily modifiable, which hinders their use for other contexts. The need for large amounts of labeled data could be partially addressed by employing unsupervised ML or GAI at the early stages of LP development to identify possible relevant patterns. However, the automatic scoring of more assessments to diagnose student performances or improve AI model accuracy at specific LP levels will eventually require a large amount of labeled data for training purposes. This challenge is especially considerable for scoring performance assessments that measure different but closely related constructs, such as those described above. Given the time required to develop supervised ML-based models and their rigid nature, it might be challenging to use these methods for LP validation for a large range of complex constructs. Recent studies in applying deep (machine) learning commonly used in GAI such as large language models, transformers, and neural networks, have provided possibilities for addressing these challenges (Sung et al., 2021; Wulff et al., 2023). Specifically, these approaches utilize a common "base" model or library for ML, and then the model is "tuned" to the specific application (or assessment construct). This represents an avenue of future research to test if such AI-based approaches are more generalizable to these educational contexts and examine if these approaches reduce the initial effort of ML model development.

The emergence of GAI offers a promising way of scoring LP-aligned assessments by directly training a GAI algorithm to use the rubric to score a wide range of student responses without the need for previously labeled data sets. For example, preliminary results on using GPT-based AI models to score NGSS LP-aligned assessments when provided the rubric show promise. However, as mentioned above, it is important to ensure the validity of the resulting AI-based scores. A possible way of building the validity argument could be holding AI algorithms to the same standards one would hold human scorers, including testing and re-testing reliability with multiple data sets to ensure that the AI-based scores are consistently reliable and bias-free across time and multiple data sets (Kaldaras, et al., 2024). Further, additional studies are needed to investigate ways to train GAI models to evaluate student responses beyond specific LP-aligned items. A promising future research avenue should focus on leveraging LPs as a basis for prompt generation to develop generalizable GAI models capable of evaluating student performance with respect to LP levels over a wide range of LP-aligned tasks.

Design-based challenges

Designed-based challenges of using AI scores for LP validation are numerous. First, a key design challenge is developing rubrics for AI scoring that can lead to high human–computer agreement and preserve the performance nature and purpose of tasks simultaneously. Specifically, most LP-aligned performance assessments represent scenario-based tasks and are usually scored holistically directly into an LP level (Jescovitch et al., 2021; Kaldaras et al., 2021a). These holistic rubrics usually represent polytomous scales with each scoring level being mutually exclusive to others and capturing a unique set of characteristics related to the complex construct described by the LP (Jescovitch et al., 2021). In contrast, some approaches to automatic scoring short, content-based CRs commonly use analytic rubrics (Liu et al., 2014; Moharreri et al., 2014; Sieke et al., 2019). Analytic rubrics are a series of binary rubrics that identify the presence or absence of specific construct-relevant ideas within responses.

A key factor in choosing either a holistic or analytic coding approach is human inter-rater reliability (IRR), which indicates the level of agreement across multiple human scorers. Analytic coding has been shown to reduce coding complexity for humans and potentially lead to better ML model performance (Jescovitch et al., 2019; Wang et al., 2021). However, both holistic (Anderson et al., 2018; Prevost et al., 2016) and analytic approaches (Haudek et al., 2012; Sieke et al., 2019) have been used to yield well-functioning ML models for short CR items.

Prior studies comparing analytic and holistic approaches to human coding for LP-aligned assessments found that analytically coded responses showed equal or better ML model performance as compared to holistic scores (Jescovitch et al., 2021). Another study showed that analytic rubrics deconstructed from holistic rubrics demonstrated moderate to a high human–computer agreement when recombined at the holistic level (Mao et al., 2018). In previous work to validate LP-aligned CR assessments of scientific argumentation (Wilson et al., 2024), the researchers found it challenging to achieve high IRR between human scorers using holistic rubrics, especially at higher LP levels. Therefore, to simplify the coding task, analytic rubrics were developed that focused on dichotomous argumentation element scoring, which led to better IRR between human coders. However, analytic scores were not meaningful for assigning responses to LP levels. The researchers then decided to recombine the analytic scores to yield a single holistic score, aligned to an LP level, for each response on an item. One caveat of this approach is that analytic rubrics should be designed to avoid scoring content and practices separately since relevant practices and content develop together (Kaldaras et al., 2022b). This will ensure that the performance nature of the assessments is preserved and that ML algorithms produce meaningful and valid scores in the context of relevant LPs. From that perspective, recent studies demonstrated a process for designing analytic rubrics that preserve the 3D nature of NGSS LP-aligned assessments (Kaldaras et al., 2022b, 2024). Another recent study also found that meaningfully designed analytics rubrics made it easier to diagnose the nature of the inaccurate scores on LP-aligned assessments, therefore contributing to improved validity of the resulting ML-based scores (Kaldaras & Haudek, 2022a). The second design-based challenge relates to the discussion of unsupervised ML. It relates to ensuring the validity of AI scores with respect to the construct of interest. In the context of supervised ML, we need to ensure that the validity argument goes beyond developing analytic scoring rubrics that yield good human–machine agreement as the sole indicator of construct validity. It is important to ensure that the resulting ML scores in each rubric category reflect the same aspects of the LP as intended by the assessment developers and the trained human scorers.

Validity challenges

Traditionally, performance metrics of how closely the machine scores match the human scores provide validity evidence for the automated scoring process (Williamson et al. 2012), and often, once the ML scoring models meet acceptable benchmarks, the ML model is used to score any number of new responses. In the context of validating LP-aligned performance assessments, this is not enough to ensure the validity of the automatic scores. Rather, the researchers need to demonstrate that an assessment probes the targeted LP levels and that the machine is scoring the same aspects of the LPs as intended by the assessment developers and the human scorers. For example, employing latent variable modeling to compare true scores for human and machine generated scores would provide additional validity evidence for the correspondence between human and ML-based scores beyond observed scores (Kaldaras & Haudek, 2022a).

Further, even if the validity of ML-based scores in relation to LP levels has been demonstrated, it is important to ensure that scores for individual students are accurate and reliable. This will allow for accurate LP-level placement of each student, opening numerous possibilities for using LPs to guide instruction and inform the development of automatic assessment systems for formative and summative uses. Depending on the assessment purpose, the automatic assessment systems can produce automatic and LP-aligned feedback targeted at individual students,

teachers, or administrators. A more detailed discussion of assessment validity in the context of GAI-based assessments has been previously published (Kaldaras et al., 2024).

Stage 3: providing LP-aligned feedback

Promises and challenges of leveraging AI

One of the central values of LP-based instruction stems from the potential of LPs to adjust the learning process to the needs of individual diverse learners by providing targeted and cognitively appropriate LP-aligned feedback. That feedback, of course, is only helpful if it is provided in a timely fashion to all the stakeholders, including individual students and teachers. Research has been conducted using automatic ML-based feedback systems focusing on supporting students in developing written explanations and arguments in science (Lee et al., 2019; Nakamura, et al, 2016; Zhu et al., 2017), science inquiry (Gobert et al., 2013) and interacting with simulations coupled with CR assessments (Kaldaras et al., 2024a, 2024b, 2024c; Lee et al., 2021). These and other studies' results indicate that automatic scoring systems can be implemented as part of formative assessment and that students' ability to construct scientific arguments can improve when receiving automatic feedback (Lee et al., 2019; Zhu et al., 2017). These studies show the promise of using ML-based automatic feedback systems to evaluate large numbers of responses quickly and thereby help many students develop higher proficiency levels on LP-aligned constructs. However, more work needs to be done on designing automatic feedback systems aligned to LPs describing complex constructs. The automatic feedback aligned to these LPs should be focused on the underlying cognition model (i.e., specific to both content and practice) and ensure that individual students receive feedback tailored to the ideas that require elaboration or are missing from their answers. Moreover, it is important to research the type of LP-aligned cognitive feedback that diverse learners find useful under various learning circumstances for attaining higher LP levels.

One example of a project aiming to implement a complex construct LP-based vision of learning in practice focuses on designing a supervised ML-based automatic scoring system that would provide immediate NGSS LP-aligned feedback to individual students on their scientific models and evidence-based explanations (He et al., 2024; Kaldaras et al., 2024). In this project, we aim to leverage previously validated NGSS-aligned LP and the associated constructed response items and provide students with opportunities to revise their models and explanations during the 1-year curriculum following targeted LP-aligned feedback. The project aims to test the usefulness of this system in helping diverse learners attain higher LP levels as a result of the personalized supervised ML-base automatic feedback on their scientific models and explanations. One of the central challenges in this project is to design useful and actionable feedback for individual learners such that they are more likely to revise their responses by incorporating the feedback rather than remain with their initial ideas. We believe the key to designing such feedback starts with the items and the LP itself. Specifically, it is important to ensure that the LP levels are described as clear and specific performance expectations that reflect how students demonstrate their understanding at a given level. Next, the items should discriminate when measuring student performance with respect to each level. Assuming the LP and the associated assessments exhibit these properties, it is possible to design feedback statements that would target specific ideas and reasoning that students struggle with.

We have designed useful and actionable feedback statements that we plan to pilot with students to further study the types of approaches to feedback design that learners would find most helpful in revising their work. This is also an important research avenue for the intersection of supervised and GAI use in the future. Specifically, our research shows less diversity in student responses at higher LP levels. Supervised ML approaches are effective in both scoring and providing pre-defined feedback to these responses. At lower LP levels, however, it becomes more challenging to reliably score student responses as they exhibit a wide range of ways of thinking. GAI provides a promising way of scoring these responses. For example, recent work shows that instructors and students find GAI-generated feedback in the context of a Physics course useful and meaningful, although the items were not aligned to an LP (Wan & Chen, 2024). However, additional work is needed to evaluate the effectiveness of GAI-based feedback when students use non-standard language and terminology, which is often the case at lower LP levels. Further, the researchers need to conduct more work on investigating ways to train GAI to provide concise, meaningful, and actionable LP-aligned feedback. Preliminary findings indicate that the feedback provided by generative AI is usually long and does not directly address LP-related ideas in student responses. Developing strategies for training GAI to provide meaningful and actionable LP-aligned feedback is an important future research avenue.

Stage 4: supporting positive learning outcomes based on the feedback

Promises and challenges of leveraging AI

The final stage of using an LP focuses on designing associated curriculum materials and professional development programs focused on supporting students to

advance along the LP (Krajcik & Shin 2022). At this stage, it is important to ensure the consequential validity of the LP- that is, evaluating how well the LP supports student learning (Alonso & Elby, 2019). To tackle this issue, the LP development process should involve formulating a theory of action that describes how the LP should be used (Alonso & Elby, 2019). This can be achieved by providing timely, targeted, and accurate feedback about student progress to students and teachers. ML approaches can help achieve this by providing automatic feedback to both stakeholders. Student-facing feedback was discussed above. Teacher-facing feedback is also critical for transforming instruction and supporting students in moving up the LP levels. As a field, we know very little about how to provide meaningful and actionable teacher-facing LP-aligned feedback to facilitate positive learning outcomes. Prior studies have shown that LP-aligned feedback improved teacher formative assessment practices and helped teachers attain a more nuanced sense of student understanding, which helped with revising activities and adjusting instruction (Zhai et al., 2018). Further, automated feedback helped teachers attend to students' struggles with content and reasoning in a timely fashion (Gerard & Linn, 2016). LPs also help teachers develop actionable pathways for adjusting instruction (Alonzo & Elby, 2019).

The key challenges in developing LP-aligned teacher-facing automatic feedback are ensuring that the feedback is specific to a student or group of students and that the teacher can immediately act upon the feedback. Further, the design of teacher-facing feedback should include curricular or instructional support that students need to progress toward higher LP levels. Teachers need support in understanding the results of automatic scoring systems and in using the automated LP-aligned feedback to inform their instruction. In the context of LP-aligned performance assessments, targeted professional learning programs are needed to improve assessment literacy among practitioners. Topics that require attention relate to current cognition theories, including the developmental nature of student understanding, situated cognition, and how these theoretical constructs relate to teaching in the context of LP-supported instruction. Further, in the project discussed above focusing on the design of an ML-driven NGSS LP-aligned automatic feedback system (Kaldaras et al., 2024a, 2024b, 2024c), we have discovered that professional learning (PL) focused on helping teachers understand the usefulness of an LP for driving instruction grounded in their student work is critical. Specifically, the PL sessions focus on sharing sample student responses to NGSS-aligned items and discussing how the LP framework can help move their students up the LP levels by adapting the curriculum is especially useful and helping teachers implement LP-based vision in practice. In the future, we will investigate how working with teachers on combining their knowledge of LP and AI feedback can help them guide and adapt their instruction. For example, recent research shows that combining instructor and GAI-generated feedback significantly enhances traditional teaching methods and results in more dynamic and responsive learning environments (Pahi et al., 2024). Investigating how to integrate human and GAI-provided feedback in the context of LP-based instruction, therefore represents a promising future research avenue.

Supporting multi-modal LP-aligned evaluation of student progress

Evaluating student performance under different modalities is essential for creating equitable learning environments. For example, supporting culturally and linguistically diverse students in building modeling skills provides an alternative mode of communicating their understanding, which is essential for equitable science assessment and engagement (Grapin & Lee, 2022). Consequently, we need to work towards developing AI-based models capable of evaluating student LP-aligned performance on multi-modal assessments. This calls for the need to go beyond text-based responses. Specifically, most of the work in automated scoring of STEM assessments has focused on CR items requiring short text responses probing the practices of argumentation (Lee et al., 2019, 2021; Zhu et al., 2017), inquiry (Gobert, et al, 2013; Linn et al., 2014) or constructing explanation (Kaldaras, Yoshida, Haudek, 2022), all of which are grounded in student language. However, to transform science education, automatic scoring approaches to a wide range of content and practices beyond text responses need development. For example, a central scientific practice that short text-based CR items cannot fully assess is scientific modeling (Schwarz et al., 2017). To demonstrate proficiency in scientific modeling, students usually need to develop (draw) a representation that provides a causal account of the phenomenon or a proposed solution to the problem. In this context, modeling closely relates to specific disciplinary content knowledge reflecting complex constructs describing usable knowledge. While other aspects of the modeling practice are important (such as developing a mathematical representation of the observed phenomenon), in the context of ML-based image recognition, we focus on drawn models as they are explicitly emphasized as an important aspect of K-12 learning (NRC, 2012).

Currently, there is limited research available on automated scoring of student models, although deep (machine) learning is used to make predictions on a

variety of data types (Kaldaras et al., 2024a, 2024b, 2024c; LeCun et al., 2015; Lee et al., 2023). The major challenge of the validity of machine scores also exists for the automatic scoring of drawn models. In the context of automated scoring of models, special attention is needed to ensure that the machine scores reflect the modeling skills as opposed to the student's ability to develop representations with artistic elements that don't necessarily relate to the practice of modeling (Leong et al., 2018). This issue could be addressed at several stages: both at the stage of the development and validation of a LP targeting the practice of modeling and at the stage of developing AI-based systems and validating ML-based scores. At the stage of developing and validating the LP, attention should be drawn to how learners demonstrate modeling proficiency in practice, what elements should be present in student responses that would serve as evidence of proficiency in modeling, and not other unrelated skills. At the stage of AI model development, attention should be paid to examining AI-based scores to ensure that the score assignment is consistent for both simple and sophisticated representations of models at all LP levels. It is also important to ensure that the AI algorithm is evaluating the same model components that a human scorer does when assigning a score (Sagherian et al., 2022). This is a critical property of an AI model because it will help build a validity argument for the resulting AI-based scores.

Further, an even more complex case of multi-modality in the context of LP-aligned performance assessments relates to simultaneous evaluation and feedback on multiple modalities within the same assessment. An example of such assessment in the context of modeling would be asking students to develop a scientific model (e.g., drawing) of the phenomenon in question and provide a written explanation of their model. When evaluating such assessments, one would need to provide feedback on the representation and the written part of the task. Recent studies show that GAI-based model (GPT-4) is effective in recognizing multiple modalities across chemistry domains, including images, diagrams, hand-written calculations, and text-based explanations to guide personalized learning experiences for individual students (Alasadi & Baiz, 2024). These results show promise for potentially leveraging GAI to provide personalized feedback simultaneously on multiple modalities, therefore supporting students in developing multi-modal knowledge-in-use.

Social and ethical implications of using AI in education

With the promise of using AI in education comes the responsibility of ensuring that AI does not perpetuate educational inequalities and injustices often ingrained in datasets (however small or big) that are used to train AI algorithms for different purposes (Akgun & Krajcik, 2024). In the context of using AI to help develop, validate, and implement LPs, it is essential to learn to recognize and minimize or eliminate the potential biases and validity threats at each of those stages. Barnes et al. (2024) provide several possible tools to support the ethical use of AI over all stages of educational research endeavors. In using AI to guide LP-based instruction, it is essential to ensure that the diversity of ways of thinking and knowing is represented in samples used to train the AI models and recognized and accounted for at the stage of using AI to guide this process. Researchers should be mindful of potential biases and how they can affect the learning experiences of students from various backgrounds. Future research should go beyond the issue of bias and tackle broader issues of assessment validity in the context of using AI-based outcomes for various purposes in education (Kaldaras, Akaeze, Reckase, 2024). There is considerable research needed on designing and testing appropriate ways of validating AI-based information on student performance and beyond for various intended uses in education. Only once a compelling validation approach is in place and sufficient validity evidence is presented should an AI algorithm be deemed suitable for use in specific educational settings.

Furthermore, it is vital to ensure that the AI-driven LP-based instructional supports provided to teachers consider the context and provide a range of culturally appropriate suggestions for a student group. This calls for researchers to pay special attention to culturally relevant information when training the AI. For example, Suresh et al., (2021, June) discuss training AI to help guide equitable classroom discussions in mathematics classrooms. In the context of LP-drive discussion, accounting for diverse ways of thinking at the LP validation stage can help obtain more equitable and culturally relevant AI-guided outcomes. Further, combining human and AI-based guidance during the learning process could help improve equity and support learning opportunities among diverse student populations (Chine et al., 2022, July). These and other research avenues should be further explored to ensure equitable and culturally relevant implementation of AI-based LP learning systems.

Conclusion

This manuscript highlights how AI approaches can assist in developing, validating, and implementing a LP-based vision of education focused on supporting learners in developing proficiency in complex constructs combining elements of disciplinary knowledge and practice using performance-based CR assessments. Employing unsupervised and semi-supervised

ML and GAI approaches during the early stages of LP development may help identify student response patterns, which reflect how proficiency in a construct develops. Further employing supervised ML and GAI at later stages of LP development and LP-based student response evaluation can streamline validation and evaluation of LP-aligned performance assessments aimed at diagnosing group and individual LP level placement with a high degree of accuracy. Similarly, AI holds promise for streamlining the process of delivering LP-aligned feedback to teachers and learners with the purpose of achieving positive learning outcomes and adjusting the learning process to the needs of individual diverse learners. However, despite significant promise, there exist considerable challenges to using AI technology at all stages of LP development and use, specifically issues related to the heterogeneity of student reasoning with complex constructs that combine elements and disciplinary knowledge and practices as well as validity and bias issues. GAI holds promise in alleviating some of the challenges of using supervised ML. We have discussed these points in detail and outlined future research avenues to tackle these issues. We hope that this short paper can serve as a guide for researchers and educators interested in studying ways of leveraging AI for implementing a LP-based vision of education into practice.

Abbreviations

LP	Learning progression
SEP	Scientific and engineering practice
DCI	Disciplinary core idea
CCC	Crosscutting concept
ML	Machine learning
NLP	Natural language processing
AI	Artificial intelligence
CR	Constructed response
MC	Multiple choice
STEM	Science, technology, engineering and math

Acknowledgements
We would like the editors and the reviewers for providing helpful comments throughout the review process. All the feedback contributed to improving the manuscript.

Author contributions
LK, KH and JK developed the outline for the paper. LK wrote the first version of the paper. KH and JK reviewed, commented, and helped revise the paper in its current form.

Funding
This work is funded by the NSF Grant # 2200757.

Data availability
There is no data associated with this manuscript.

Declarations

Competing interests
The authors have no competing interests to disclose.

Received: 7 June 2023 Accepted: 15 October 2024
Published online: 08 November 2024

References

Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex physics problems. *American Journal of Physics, 83*(5), 459–467.

Akgun, S., & Krajcik, J. (2024). Artificial intelligence (AI) as the growing actor in education: raising critical consciousness towards power and ethics of AI in K-12 STEM classrooms. In X. Zhai & J. Krajcik (Eds.), *Uses of artificial intelligence in STEM education*. Oxford: Oxford University Press.

Alasadi, E. A., & Baiz, C. R. (2024). Multimodal generative artificial intelligence tackles visual problems in chemistry. *Journal of Chemical Education, 101*(7), 2716–2729.

Alonzo, A. C., & Elby, A. (2019). Beyond empirical adequacy: Learning progressions as models and their value for teachers. *Cognition and Instruction, 37*(1), 1–37.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*(3), 389–421.

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock, J. B., & Welch, M. M. (2018). Designing educational systems to support enactment of the Next Generation Science Standards. *Journal of Research in Science Teaching, 55*(7), 1026–1052.

Barnes, T., Danish, J., Finkelstein, S., Molvig, O., Burriss, S., Humburg, M., Reichert, H., Limke, A. (2024). Toward Ethical and Just AI in Education Research. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

Berland, M., Martin, T., Benton, T., Smith, C. P., & Davis, D. (2013). Using learning analytics to understand the learning pathways of novice programmers. *The Journal of the Learning Sciences, 22*(4), 564–599.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn* (Vol. 11). National academy press.

Brown, N. J., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review, 23*(2), 221.

Burkholder, E., Blackmon, L., & Wieman, C. (2020). Characterizing the mathematical problem-solving strategies of transitioning novice physics students. *Physical Review Physics Education Research, 16*(2), 020134.

Burkholder, E., Hwang, L., & Wieman, C. (2021). Evaluating the problem-solving skills of graduating chemical engineering students. *Education for Chemical Engineers, 34*, 68–77.

Butterfuss, R., and Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. Paper presented at generative artificial intelligence for measurement and education meeting. https://hdoran.github.io/Blog/ContentMapping.pdf

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating mixed methods research with natural language processing of big text data. *Journal of Mixed Methods Research, 15*(3), 398–412. https://doi.org/10.1177/15586898211021196

Chine, D. R., Brentley, C., Thomas-Browne, C., Richey, J. E., Gul, A., Carvalho, P. F., & Koedinger, K. R. (2022). Educational equity through combined human-AI personalization: A propensity matching evaluation. In D. Chine (Ed.), *International Conference on Artificial Intelligence in Education* (pp. 366–377). Springer International Publishing.

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching: THe Official Journal of the National Association for Research in Science Teaching, 46*(6), 606–609.

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th–10th grades. *Journal of Research in Science Teaching: THe Official Journal of the National Association for Research in Science Teaching, 46*(6), 655–674.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. (2007). *Taking science to school: Learning and teaching science in grades K-8*. National Academy Press.

Eignor, D. R., et al. (2013). The standards for educational and psychological testing" in APA handbook of testing and assessment in psychology. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I.C. Hansen, N. R. Kuncel, & S. P. Reise (Eds.), *Test theory and testing and assessment in industrial and organizational psychology* (Vol. 1, pp. 245–250). American Psychological Association.

Finnish National Board of Education (FNBE). (2015). National core curriculum for general upper secondary schools 2015. Helsinki, Finland: Finnish National Board of Education (FNBE). http://www.oph.fi/saadokset_ja_ohjeet/opetussuunnitelmien_ja_tutkintojen_perusteet/lukiokoulutus/lops2016/103/0/lukion_opetussuunnitelman_perusteet_2015

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education, 27*(1), 111–129.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43–57.

Gobert, J. D., Pedro, M. S., Raziuddin, J., & Baker, R. S. (2013). From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining. *The Journal of the Learning Sciences, 22*(4), 521–563.

Gorin, J. S., & Mislevy, R. J. (2013, September). Inherent measurement challenges in the next generation science standards for both formative and summative assessment. In Invitational research symposium on science assessment.

Grapin, S. E., & Lee, O. (2022). WIDA English language development standards framework, 2020 edition: Key shifts and emerging tensions. *TESOL Quarterly, 56*(2), 827–839.

Gunckel, K. L., Covitt, B. A., Berkowitz, A. R., Caplan, B., & Moore, J. C. (2022). Computational thinking for using models of water flow in environmental systems: Intertwining three dimensions in a learning progression. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21755

Hammer, D., & Sikorski, T. R. (2015). Implications of complexity for research on learning progressions. *Science Education, 99*(3), 424–431.

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.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Life Sciences Education, 11*(3), 283–293.

He, P., Shin, N., Kaldaras, L., & Krajcik, J. (2024). Integrating artificial intelligence into learning progression to support student knowledge-in-use: Opportunities and challenges. In H. Jin, D. Yan, & J. Krajcik (Eds.), *Handbook of research on science learning progressions* (pp. 461–487). New York: Routledge.

Herrmann-Abell, C. F., & DeBoer, G. E. (2018). Investigating a learning progression for energy ideas from upper elementary through high school. *Journal of Research in Science Teaching, 55*(1), 68–93.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., & Haudek, K. C. (2019). Deconstruction of holistic rubrics into analytic rubrics for large-scale assessments of students' reasoning of complex science concepts. *Practical Assessment, Research, and Evaluation, 24*(1), 7.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). 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, 30*(2), 150–167.

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned Next Generation Science Standards performance assessments. *Front. Educ., 7*, 968289. https://doi.org/10.3389/feduc.2022.968289

Kaldaras, L. (2020). *Developing and validating NGSS-aligned 3D learning progression for electrical interactions in the context of 9th grade physical science curriculum*. Michigan State University.

Kaldaras, L., & Krajcik, J. (2024). Development and validation of knowledge-in-use learning progressions. In H. Jin, D. Yan, & J. Krajcik (Eds.), *Handbook of research on science learning progressions* (pp. 70–87). New York: Routledge.

Kaldaras, L., & Wieman, C. (2023a). Cognitive framework for blended mathematical sensemaking in science. *International Journal of STEM Education, 10*(1), 18.

Kaldaras, L., & Wieman, C. (2023b). Instructional model for teaching blended math-science sensemaking in undergraduate science, technology, engineering, and math courses using computer simulations. *Physical Review Physics Education Research, 19*(2), 020136.

Kaldaras, L., Akaeze, H. O., & Krajcik, J. (2023). Developing and validating an Next Generation Science Standards-aligned construct map for chemical bonding from the energy and force perspective. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21906

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. In L. Kaldaras (Ed.), *Frontiers in education* (Vol. 9, p. 1399377). Frontiers Media SA.

Kaldaras, L., Akaeze, H., & Krajcik, J. (2021a). 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, 58*(4), 589–618.

Kaldaras, L., Akaeze, H., & Krajcik, J. (2021b). A methodology for determining and validating latent factor dimensionality of complex multi-factor science constructs measuring knowledge-in-use. *Educational Assessment, 26*, 1–23.

Kaldaras, L., Li, T., Haudek, K., & Krajcik, J. (2024b). *Developing rubrics for AI scoring of NGSS learning progression-based scientific models* (p. 2024). Paper presented at American Educational Research Association.

Kaldaras, L., Wang, K. D., Nardo, J. E., Price, A., Perkins, K., Wieman, C., & Salehi, S. (2024c). Employing technology-enhanced feedback and scaffolding to support the development of deep science understanding using computer simulations. *International Journal of STEM Education, 11*(1), 30.

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. In *Frontiers in education* (Vol. 7, p. 983055). Frontiers Media SA.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*, 249–268.

Krajcik, J. S. (2021). Commentary—Applying Machine Learning in Science Assessment: Opportunity and Challenges. *Journal of Science Education and Technology, 30*(2), 313–318.

Krajcik, J. S., Namsoo. (2023). Student concepts, conceptual change and learning progressions. In: N. G. Lederman, Zeidler, D.L, Lederman, J.S. (Eds). Handbook of Research on Science Education. Taylor and Francis group

Kubsch, M., Krist, C., & Rosenberg, J. M. (2022). Distributing epistemic functions and tasks—a framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21803

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.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

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, 30*(2), 168–192.

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education, 103*(3), 590–622.

Lee, J., Lee, G. G., & Hong, H. G. (2023). Automated assessment of student hand drawings in free-response items on the particulate nature of matter. *Journal of Science Education and Technology, 32*(4), 549–566.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Education design and cognitive science* (Vol. 5, pp. 101–169). Lawrence Erlbaum Associates.

Leong, C. W., Liu, L., Ubale, R., & Chen, L. (2018, June). Toward large-scale automated scoring of scientific visual models. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1–4).

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science, 344*(6180), 155–156. https://doi.org/10.1126/science.1245980

Litman, D. (2016). Natural language processing for enhancing teaching and learning. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 4170–4176.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*(2), 19–28. https://doi.org/10.1111/emip.12028

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment, 23*(2), 121–138.

Martin, P. P., & Graulich, N. (2024). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence, 7*, 100265.

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21903

Mccomas, W. (2014). *The atlas of science literacy*. SensePublishers. https://doi.org/10.1007/978-94-6209-497-0_8

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012.

Ministry of Education, P. R. China. (2020). *Curriculum plan for senior high school [普通高中课程方案]*. People's Education Press.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416.

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

Mohan, L., Chen, J., & Anderson, W. A. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*(6), 675–698.

Moharreri, K. M., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach, 7*, 15.

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., & Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research, 12*(1), 010122. https://doi.org/10.1103/PhysRevPhysEducRes.12.010122

National Academies of Sciences, Engineering, and Medicine. (2019). *Science and engineering for grades 6–12: Investigation and design at the center*. The National Academies Press.

National Research Council. (2006). *Systems for state science assessment*. The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. The National Academies Press.

Nelson, L. K. (2020). Computational grounded theory: a methodological framework. *Sociological Methods & Research, 49*(1), 3–42. https://doi.org/10.1177/0049124117729703

News and events. National Assessment Governing Board Approves an Updated Science Framework for the 2028 Nation's Report Card. (2023, November 17). https://www.nagb.gov/news-and-events/news-releases/2023/updated-science-framework-2028.html

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

Pahi, K., Hawlader, S., Hicks, E., Zaman, A., & Phan, V. (2024). Enhancing active learning through collaboration between human teachers and generative AI. *Computers and Education Open, 6*, 100183.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press, 2102 Constitutions Avenue, NW, Lockbox 285, Washington, DC 20055.

PISA 2025 Science Framework Draft. https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf .

Prevost, L. B., Smith, M. K., & Knight, J. K. (2016). Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. *CBE—Life Sciences Education, 15*(4), ar65.

Price, A., Salehi, S., Burkholder, E., Kim, C., Isava, V., Flynn, M., & Wieman, C. (2022). An accurate and practical method for assessing science and engineering problem-solving expertise. *International Journal of Science Education, 44*(13), 2061–2084.

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*. https://doi.org/10.1007/s10956-020-09862-4

Sagherian, A., Lingaiah, S., Abouelenien, M., Leong, C. W., Liu, L., Zhao, M., Lafuente, B., Chen, S.-K., & Qi, Y. (2022). Learning Progression-based Automated Scoring of Visual Models. *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 213–222. https://doi.org/10.1145/3529190.3529192

Samala, A. D., Zhai, X., Aoki, K., Bojic, L., & Zikic, S. (2024). An in-depth review of ChatGPT's pros and cons for learning and teaching in education. *International Journal of Interactive Mobile Technologies, 18*, 96–117. https://doi.org/10.3991/ijim.v18i02.46509

Sao Pedro, M. A., Gobert, J. D., & Betts, C. G. (2014). Towards scalable assessment of performance-based skills: Generalizing a detector of systematic science inquiry to a simulation with a complex structure. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 591–600). Springer International Publishing.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world using next generation science and engineering practices*. NSTA Press.

Scott, E. E., Cerchiara, J., McFarland, J. L., Wenderoth, M. P., & Doherty, J. H. (2022). How students reason about matter flows and accumulations in complex biological phenomena: An emerging learning progression for mass balance. *Journal of Research in Science Teaching, 60*, 63.

Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In R. J. Shavelson (Ed.), *Learning progressions in science* (pp. 13–26). Brill Sense.

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences, 22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sieke, S. A., McIntosh, B. B., Steele, M. M., & Knight, J. K. (2019). Characterizing students' ideas about the effects of a mutation in a noncoding region of DNA. *CBE-Life Sciences Education, 18*(2), ar18. https://doi.org/10.1187/cbe.18-09-0173

Sikorski, T. R. (2019). Context-dependent "upper anchors" for learning progressions. *Science & Education, 28*(8), 957–981.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). FOCUS ARTICLE: Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective, 4*(1–2), 1–98.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*, 610–631.

Sripathi, K. N., Moscarella, R. A., Steele, M., Yoho, R., You, H., Prevost, L. B., Urban-Lurain, M., Merrill, J., & Haudek, K. C. (2023). Machine learning mixed methods text analysis: An illustration from automated scoring models of student writing in biology education. *Journal of Mixed Methods Research*. https://doi.org/10.1177/15586898231153946

Sung, S. H., Li, C., Chen, G., et al. (2021). How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology, 30*, 210–226. https://doi.org/10.1007/s10956-020-09856-

Suresh, A., Jacobs, J., Clevenger, C., Lai, V., Tan, C., Martin, J. H., & Sumner, T. (2021). Using ai to promote equitable classroom discussions: The talk moves application. In A. Suresh (Ed.), *International conference on*

artificial intelligence in education (pp. 344–348). Springer International Publishing.
Tschisgale, P., Wulff, P., & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research, 19*(2), 020123. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123
Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research, 20*(1), 010152.
Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated Scoring of Chinese Grades 7–9 Students' Competence in Interpreting and Arguing from Evidence. *Journal of Science Education and Technology, 30*(2), 269–282. https://doi.org/10.1007/s10956-020-09859-z
Wiley, J., Hastings, P., Blaum, D., et al. (2017). Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education, 27*, 758–790. https://doi.org/10.1007/s40593-017-0138-z
Willard, T. (2020). *The NSTA Atlas of the Three Dimensions*. NSTA Press. 1840 Wilson Boulevard, Arlington, VA 22201.
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x
Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., & Zhai, X. (2024). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21864
Wind, S. A., Alemdar, M., Lingle, J. A., Moore, R., & Asilkalkan, A. (2019). Exploring student understanding of the engineering design process using distractor analysis. *International Journal of STEM Education, 6*(1), 1–18.
Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412.
Yao, J. X., & Guo, Y. Y. (2018). Core competences and scientific literacy: The recent reform of the school science curriculum in China. *International Journal of Science Education, 40*(15), 1913–1933.
Zaballa, O., Pérez, A., Inhiesto, E. G., Ayesta, T. A., & Lozano, J. A. (2023). Learning the progression patterns of treatments using a probabilistic generative model. *Journal of Biomedical Informatics, 137*, 104271.
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.
Zhu, M., Lee, H. S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education, 39*(12), 1648–1668.

Publisher's Note