frontiers | Frontiers in Education

TYPE Conceptual Analysis
PUBLISHED 07 August 2024
DOI 10.3389/feduc.2024.1399377

OPEN ACCESS

EDITED BY
Janet Clinton,
The University of Melbourne, Australia

REVIEWED BY
Yan Liu,
Carleton University, Canada
Ahmed Kharrufa,
Newcastle University, United Kingdom

*CORRESPONDENCE
Leonora Kaldaras
✉ leonora.kaldaras@colorado.edu;
✉ kaldaras@stanford.edu;
✉ lkaldara@ttu.edu

RECEIVED 13 March 2024
ACCEPTED 17 July 2024
PUBLISHED 07 August 2024

CITATION
Kaldaras L, Akaeze HO and Reckase MD (2024) Developing valid assessments in the era of generative artificial intelligence.
Front. Educ. 9:1399377.
doi: 10.3389/feduc.2024.1399377

Developing valid assessments in the era of generative artificial intelligence

Leonora Kaldaras[1,2]*, Hope O. Akaeze[3,4] and Mark D. Reckase[3]

[1]Department of Physics, University of Colorado Boulder, Boulder, CO, United States, [2]Graduate School of Education, Stanford University, Stanford, CA, United States, [3]Michigan State University, East Lansing, MI, United States, [4]Community Evaluation Programs, Office of Public Engagement and Scholarship, University Outreach and Engagement, Michigan State University, East Lansing, MI, United States

Generative Artificial Intelligence (GAI) holds tremendous potential to transform the field of education because GAI models can consider context and therefore can be trained to deliver quick and meaningful evaluation of student learning outcomes. However, current versions of GAI tools have considerable limitations, such as social biases often inherent in the data sets used to train the models. Moreover, the GAI revolution comes during a period of moving away from memorization-based education systems toward supporting learners in developing the ability to apply knowledge and skills to solve real-world problems and explain real-world phenomena. A challenge in using GAI tools for scoring assessments aimed at fostering knowledge application is ensuring that these algorithms are scoring the same construct attributes (e.g., knowledge and skills) as a trained human scorer would score when evaluating student performance. Similarly, if using GAI tools to develop assessments, one needs to ensure that the goals of GAI-generated assessments are aligned with the vision and performance expectations of the learning environments for which these assessments are developed. Currently, no guidelines have been identified for assessing the validity of AI-based assessments and assessment results. This paper represents a conceptual analysis of issues related to developing and validating GAI-based assessments and assessment results to guide the learning process. Our primary focus is to investigate how to meaningfully leverage capabilities of GAI for developing assessments. We propose ways to evaluate the validity evidence of GAI-produced assessments and assessment scores based on existing validation approaches. We discuss future research avenues aimed at establishing guidelines and methodologies for assessing the validity of AI-based assessments and assessment results. We ground our discussion in the theory of validity outlined in the Standards for Educational and Psychological Testing by the American Educational Research Association and discuss how we envision building on the standards for establishing the validity of inferences made from the test scores in the context of GAI-based assessments.

KEYWORDS

generative artificial intelligence (GAI), validity, knowledge application, validity standards, assessment practices, evaluation of cognitive development with GAI

Introduction

Recent advances in natural language processing (NLP) and deep learning technology have led to the development of models that can process language and perform a wide range of tasks such as generating high-quality text, images, and other content. GAI holds tremendous potential to transform education because the GAI models can be trained to perform specific tasks and potentially automate or streamline various processes. These models are pre-trained on large volumes of data and are broadly referred to as Generative Artificial Intelligence (GAI) tools. For example, one of the most popular GAI 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 (Samala et al., 2024). However, current GAI tools have considerable limitations, such as social biases often inherent in the data sets used to train these models (Mao et al., 2024). These biases, among other factors, must be considered when implementing GAI tools in education.

Moreover, the GAI revolution comes during a period of significant changes in global education. Specifically, recent educational reforms worldwide emphasize supporting learners in developing the ability to apply knowledge to solve real-world problems and explain real-world phenomena. Examples include PISA, which has emphasized knowledge application on their assessments (OECD, 2016). Further, Germany (Kulgemeyer and Schecker, 2014) and Finland [Finnish National Board of Education (FNBE), 2015] have developed national standards focused on supporting learners in developing and measuring competencies. Competencies refer to standards expressed as learning goals requiring learners to apply their knowledge rather than reciting memorized information. A similar push toward measuring competencies occurs in the Chinese educational system (Ministry of Education, P. R. China, 2018; Yao and Guo, 2018). In the United States, similar efforts have resulted in the publication of the Framework for K-12 Science Education (*the Framework*) and the Next Generation Science Standards (NGSS), which emphasize fostering knowledge growth coherently over time so learners can apply what they learn (National Research Council, 2012; NGSS Lead States, 2013). The National Assessment Governing Board (NAGB) has just 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 move toward supporting knowledge application skills calls for significant changes at all stages of the learning process, including the development of new learning systems that help foster knowledge application. The learning system includes curriculum and assessment materials and the necessary instructional support. Effective development and implementation of such learning systems depends on the ability to adjust the learning process to the needs of academically, culturally, and linguistically diverse learners. Specifically, supporting the development of complex understanding related to knowledge application requires that students have opportunities to learn the appropriate knowledge and skills over time. The learning process must also be appropriately scaffolded to meet the needs of individual diverse learners (National Research Council, 2012). Therefore, to effectively support knowledge application, a learning system must incorporate features such as creating meaningful learning opportunities and providing timely and appropriate scaffolding and feedback.

Consequently, teachers and learners need access to timely, informative, high-quality feedback to effectively engage in the learning process and develop knowledge application skills (Pellegrino et al., 2001; Krajcik, 2021). This will ensure that teachers can meaningfully adjust their instruction and create the necessary learning opportunities for students, and students can use this feedback to engage in discussions and self-reflection to deepen and improve their understanding. To provide this type of feedback, the assessments must effectively measure complex understanding, particularly the ability to apply relevant knowledge and skills. This, in turn, calls for moving away from multiple-choice (MC) based assessments toward open-ended assessments that require students to engage in developing models and explanations of phenomena (Krajcik, 2021). These assessments will allow us to measure complex reasoning and skills that are reflective of knowledge application ability and gain the information necessary for providing informative feedback to students and teachers. In short, the shift toward fostering knowledge application requires moving toward assessments that can guide the learning process instead of delivering point measurement results on the amount of information learners retained within a given time frame. Such open-ended assessments are time-consuming to develop, score, and report. AI tools, including GAI, have the potential to help tackle this challenge (Krajcik, 2021; Kaldaras et al., 2022). However, leveraging GAI tools to develop such assessments and provide feedback requires that the assessments developed using a GAI and the results of GAI model analysis of the assessment data meaningfully relate to the underlying constructs. They must offer valid and reliable measures capable of meaningfully guiding the learning process (Kaldaras and Haudek, 2022).

A construct refers to an unobservable and possibly hypothetical entity or concept. We evaluate or infer students' grasp of a construct based on their performance on assessment questions designed to measure it. Evaluation of the degree to which GAI provides an accurate evaluation of student progress on constructs describing knowledge application skills calls for careful evaluation of the validity of inferences drawn from AI-based assessment results (Kaldaras and Haudek, 2022). Validity refers to the degree to which evidence and theory support interpretations of assessment results (for example, test scores) for the proposed uses of a given assessment (Messick, 1980; Eignor, 2013). Construct validity relates to how well an assessment instrument represents and reflects the construct of interest. The validation process involves accumulating multiple relevant evidence sources to provide sound scientific support for the proposed interpretation of the assessment results (Eignor, 2013). Consequently, when assessment results are interpreted in multiple ways—for example, as a summative measure of what students have mastered or as a predictive measure of future performance—each of these intended uses of the assessment result must have evidence to support the desired inference.

In evaluating the validity of AI-based assessment outputs, an intended use of the assessment results could be to deliver timely and informative feedback to teachers and students to guide the learning process. For instance, AI-based assessments can be used to guide the learning process when they accurately diagnose students' understanding of a construct that describes knowledge application in a given context and provide meaningful feedback to teachers and students, where such feedback is focused on supporting the learners in transitioning to a higher understanding of that construct. Each of these intended uses incorporates multiple specific purposes as well. For example, teacher-facing feedback might be focused on delivering information about a student's current level of understanding or providing guidance on creating learning opportunities to compensate for the lack of prior knowledge. Student-facing feedback might be focused on supporting individual student learning through self-reflection and revision of the answer. All these intended uses of the AI-based assessment results need supporting evidence.

One of the central challenges in using GAI tools for scoring assessments that will be used to guide the learning process aimed at fostering knowledge application is ensuring that these algorithms are scoring the same construct attributes (e.g., knowledge and skills) as a trained human scorer would score when evaluating student performance on the assessment. Similarly, when using GAI to assist in developing assessments focused on evaluating knowledge application, it is critical to ensure that the GAI-generated assessments represent a valid measure of the relevant knowledge application constructs. Currently, no guidelines have been established for assessing the validity of GAI-based assessments and assessment results. The purpose of this paper is to propose ways to evaluate the validity evidence of GAI-based assessments and scores based on existing approaches and discuss future research avenues for establishing guidelines and methodologies for assessing the validity of GAI-based assessment results.

This paper represents a conceptual analysis of issues related to validating AI-based assessments and assessment results to guide the learning process. We ground our discussion in the theory of validity outlined in the Standards for Educational and Psychological Testing by the American Educational Research Association (Eignor, 2013) and discuss how we envision building on the standards for establishing the validity of inferences made from the test scores in the context of GAI-based assessments.

Structure of the paper

The paper begins with a brief historical overview of using AI approaches in education. We focus specifically on AI-based evaluation of student responses to assessments since this has been the most widespread way of using AI in education in the past. Further, we will discuss expanding use of AI to one the most critical aspects of assessment development—defining the construct of interest and the associated proficiencies. The need for defining what proficiency in a construct looks like prior to developing assessments for measuring this construct has been discussed by various educational experts [see, for example, research on construct modeling by Brown and Wilson (2011); or research on learning progressions that measure knowledge application by Kaldaras et al. (2021a,b, 2023)]. Similar need for defining proficiencies is outlined in multiple policy documents that discuss the importance of organizing the learning process along empirically derived learning progressions (National Research Council, 2012; National Assessment Governing Board, 2023). Further, the most substantial part of the paper is dedicated to evaluating assessment results and assessments generated using GAI and the associated validity evidence. In this section we focus on the types of validity evidence outlined in the Standards for Educational and Psychological Testing (Eignor, 2013), including: (1) evidence based on test content; (2) evidence based on response process; (3) evidence based on internal structure; (4) evidence based on relation to other variables; (5) validity generalization. For each type, we propose ways of evaluating validity evidence for assessments and assessment results generated by GAI and discuss future research avenues. We conclude by discussing contributions of the conceptual analysis presented in the paper and proposing future research avenues focused on standardizing the validation process of GAI-generated assessments and assessments results.

Validity and AI-based scores: a historical perspective

Before the emergence of GAI tools, various machine learning (ML) tools were used to evaluate student performance on open-ended assessments. These ML tools were often grounded in supervised or semi-supervised ML approaches that required large sets of previously labeled data for training an ML algorithm to perform specific tasks—for example, score student responses to assessments (Zhai et al., 2020; Kaldaras et al., 2022). These traditional ML algorithms focused predominantly on analyzing and interpreting data. The validity of scores produced by these ML models was evaluated by comparing agreement between human and machine-assigned scores. Therefore, human scores have historically been used as a gold standard against which the validity of ML scores was evaluated (Zhai et al., 2020; Kaldaras et al., 2022). Very little work has been done on assessing the validity of ML-based scores beyond human-machine agreement (Kaldaras and Haudek, 2022), which can be considered a criterion-based validity measure.

Generative Artificial Intelligence models are also ML models. In contrast to traditional ML models that use supervised training approaches, GAI models do not require a pre-trained data set to perform tasks. Instead, they have already been trained on all the available data before release. For example, the current version of ChatGPT is trained on all the available data until 2023. It is also possible to conduct additional training of the GAI models to perform specific tasks, which means users may further train these models on a range of examples to help tailor GAI outputs to their desired outcomes.

Unlike traditional ML algorithms trained to interpret and analyze data, GAI models are designed to create novel, original outputs. They are, therefore, more versatile in the range of tasks they can perform. GAI models are promising for evaluating student performance because they do not require large sets of previously scored assessment data. However, we believe training on outputs previously evaluated by humans should be essential to preparing GAI algorithms to perform evaluation tasks. Evaluating validity evidence from multiple sources should also be an integral part of the training process for any GAI used in education. For example, assessing the validity of AI-based scores produced by GAI models is necessary to ensure that these scores are meaningful and can be used for the intended purposes, such as providing feedback to teachers and students. Currently, no such standards exist in the field of education. We further discuss approaches that can be used to evaluate the validity of GAI-based outputs for guiding the learning process.

Using GAI tools to help define construct proficiency levels

The purpose of any assessment is to measure a student's level of proficiency in a specific construct. Construct refers to an unobserved entity (topic, set of skills, etc.) that is of interest to be measured. The first step in designing an assessment that measures a given construct is to understand what skills and knowledge reflect proficiency in that construct (Pellegrino et al., 2001; Brown and Wilson, 2011). A construct describing knowledge application is defined by carefully specifying all the aspects of content knowledge and skills that students should demonstrate at various levels of sophistication (Kaldaras et al., 2021a). The process of specifying the skills and knowledge necessary to demonstrate proficiency often results in defining a cognitive model, such as learning progression (LP), that describes a path that learners can follow to develop a higher proficiency on a construct (Duschl and Hamilton, 2011). The main advantage of cognitive models lies in their capability to serve as a roadmap for guiding instruction and adjusting the learning process to the needs of individual learners (Duschl and Hamilton, 2011; Kaldaras and Krajcik, 2024). While cognitive models are incredibly useful, defining and validating cognitive models requires large amounts of data on student performance on assessments that measure the construct (see examples in Kaldaras et al., 2021a, 2023). Obtaining and evaluating enough data to extensively define a cognitive model for a given construct is time and resource consuming. GAI models can be leveraged to identify patterns in large sets of student responses to identify meaningful clusters of response types to help further define proficiency levels of cognitive models.

Further, GAI tools can also be used to generate example responses at varying levels of sophistication in situations where student response data are not available or limited. This capability of GAI tools to potentially streamline the process of defining and validating cognitive models for various constructs has the potential to transform the field of education. In turn, researchers working on validation will evaluate the response clusters identified by GAI and judge the relevance of the GAI-identified patterns for describing proficiency in the construct of interest. Researchers can further engage in iterative cycles to train the GAI algorithms to recognize attributes relevant to the construct of interest. This process will serve a dual purpose: validating the cognitive model and training the GAI algorithm to identify different proficiency levels. The pre-trained GAI model can be used to design assessments and scoring rubrics and evaluate student performance on the assessment with respect to proficiency levels defined by the cognitive model. We further discuss these steps for the relevant validity evidence sources.

Evaluating validity evidence sources generated using GAI

Below, we discuss how different sources of validity evidence will potentially be impacted by incorporating GAI into the process of test development and evaluation of the test results. We discuss the validity evidence sources outlined in the Standards for Educational and Psychological Testing (Eignor, 2013) including: (1) evidence based on test content; (2) evidence based on response process; (3) evidence based on internal structure; (4) evidence based on relation to other variables; (5) validity generalization. Note that each of these evidence sources is not required in all settings. Instead, support is needed for each proposition that underlies the proposed test interpretation for a specific use (Eignor, 2013). For example, a proposition that the test covers a particular topic may be supported without a proposition that a test predicts a given criterion (Eignor, 2013). However, a more complex proposition, such as the test covering a particular topic and can be used to make inferences about supporting learners in transitioning to a higher-level understanding (i.e., guide the learning process), requires evidence supporting both parts of this proposition. Suppose GAI is used to generate support for any of the validity evidence sources discussed below. These sources are used to develop the validity argument for the intended use of the test scores in a given setting. In that case, GAI-generated validity evidence sources should also be evaluated to ensure that they meaningfully represent the validity evidence needed to support desired propositions. We will further discuss possible ways of assessing GAI-assisted validity evidence sources for these purposes.

Evidence based on test content

This type of evidence relates to analyzing the relationship between the test content and the construct it is intended to measure. Obtaining evidence based on test content traditionally involves specifying the test domain that describes in detail all the aspects related to content and skills measured on a test. Next, it involves analysis of the correspondence between the test domain and the test items. This analysis can be done by researchers and expert judgment on the relationship between the test domain and test components. When designing tests that measure and guide student learning this type of evidence relates to alignment—a correspondence between the learning standards (for example, the Next Generation Science Standards) and test content. In this context, evaluating evidence based on test content involves assessing whether the test appropriately measures a set of standards. Educators actively use GAI tools to develop assessment questions for different types of constructs (Gierl and Lai, 2018).

Considering that developing test items is an expensive and time-consuming process, it is highly likely that states and other test development agencies will be using GAI tools to develop test items for measuring various constructs. GAI offers a way to streamline and lower the cost of developing tests for both formative and summative use. There are several ways to gather evidence for the alignment between the GAI-generated assessments and the test domain. For example, a recent study developed an 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 and Doran, 2024). This approach could reduce the time and resources needed to perform content mapping, an essential part of the alignment process.

Further, one might use specific information from the test development process as a basis for GAI prompt generation. For example, test developers often use an evidence-centered design (ECD) approach (Mislevy et al., 2003; Kaldaras et al., 2021a, 2023) in test design. This approach involves carefully specifying an ECD argument that consists of the claim and evidence. Claim reflects what students should be able to do with the knowledge and skills. Evidence provides details on the types of evidence that should be observed in student responses to meet the claim requirements. These evidence statements are used to design assessment questions that probe a specific claim. Defining an ECD argument involves careful consideration of the test domain to improve the alignment between the test domain and the assessment questions. Therefore, using elements from the claim and the evidence as a basis for GAI prompt generation can improve the alignment between GAI-generated assessments and the test domain. Similarly, suppose there is a cognitive model that is used to guide the test development. Then, the description of the proficiency levels can be used as a basis for GAI prompt generation to guide the development of test questions. In both cases, of course, the resulting GAI output should be evaluated by humans to judge the degree of alignment between the test domain and the GAI-generated test questions. If any misalignments are observed, they should be addressed through prompt generation. Documenting this process can serve as evidence for the validity of the test content of GAI-generated tests.

A study demonstrating the basis for this approach for automatic scoring LP-aligned scientific explanations was conducted by Kaldaras et al. (2022). The study demonstrated how a validated LP and associated ECD arguments can be used to design a rubric for AI-based scoring of LP-aligned scientific explanations that measure knowledge application. This study was conducted with supervised ML, but a similar approach can be used with GAI. Specifically, LPs and ECD arguments can be used as a basis for prompt generation for designing LP-aligned assessment items and scoring rubrics. This is a promising future research avenue considering limited research currently available on GAI-assisted assessment generation.

When there are no ECD arguments or a cognitive model available (which is often the case in classroom instruction settings), one could use previously developed test questions that have been shown to measure the test domain of interest as a basis for GAI prompt generation. One would evaluate the extent to which GAI-generated assessments parallel the sample assessment question and ensure that all the new GAI-generated aspects of the assessments meaningfully align with the targeted construct features. Providing multiple examples of test questions could result in better alignment between GAI-generated assessments and the test domain, but that claim should be further investigated.

Evidence-based on response process

Using GAI to identify response process patterns in large samples

Validity evidence based on the response process refers to evaluating whether the test takers engage in the specific cognitive processes intended to be measured by the test. For example, engaging in the process of blended math-science sensemaking (MSS) involves learners demonstrating that they are integrating the relevant math and science domains when answering the test questions (Kaldaras and Wieman, 2023). Theoretical and empirical analysis of the response process provides information about the fit between the theoretical construct and the response process engaged in by test takers. Similarly, when validating a learning progression, evidence based on the response process is evaluated to judge the degree of alignment between the theoretically proposed LP proficiency levels and the actual student responses to items designed to probe those levels. If sufficient evidence is obtained to suggest that student response data support the LP levels, one can claim that the LP-aligned assessment instrument exhibits response process-based validity (Kaldaras et al., 2023). Larger samples of student responses will provide a stronger argument for response process-based validity but are also more time-consuming and expensive to evaluate.

When working with large samples of responses, GAI can be used to help assess this validity evidence by identifying clusters of patterns in student responses. Test developers, in turn, can evaluate these patterns to see if they meaningfully relate to the cognitive processes measured by the test. For example, we are currently exploring ways to evaluate CR assessments aligned to the LP for math-science sensemaking (MSS). We are using the LP as a basis for designing prompts for ChatGPT to evaluate these assessments. We also request that GPT provide a rationale for assigning LP level for each response. Through this process we are discovering that GPT is helpful in identifying specific response patterns that are important to define and incorporate into the prompt and describe in the LP. Therefore, GPT is helping us to further define the LP levels and specify different response process types that students can demonstrate when engaging in MSS at different LP levels.

Using GAI to suggest response process patterns

When student samples are small or hard to obtain, GAI might be used to generate possible sample student responses to the test questions and provide a way to get preliminary response process-based validity evidence. In this context, careful prompt generation should ensure that GAI does not offer the ideal correct answer but generates the possible answers likely provided by the target student population. One might specify the characteristics and possible prior knowledge of the target student population to ensure that GAI has more information to make more accurate suggestions of how students might respond. For example, one might provide various information sources to the GAI model, such as student grade level, previously covered materials, and student demographics, among other factors. Then, one would investigate the types of potential responses that GAI would suggest and evaluate whether the responses represent the desired LP levels. One should be very careful to explore potential GAI-generated biases inherent in the suggested responses and always aim to check the validity of proposed inferences with an actual sample of data collected from human learners. An example of GAI-generated bias in this context might refer to GAI only suggesting responses with multiple inaccuracies or responses associated with lower proficiency level for specific student populations (for example, specific demographic groups or gender groups). This type of GAI-generated bias poses threat to validity of assessments for these student groups.

Further, failure to account for non-standard language is another example of GAI-generated bias. GAI-based responses should be carefully examined to ensure that these responses contain non-standard language because students often use non-standard language in their responses to provide an accurate account of phenomena. For example, there are multiple ways a student response can reflect understanding of proportional relationships. A student might say that two variables are proportional because they change in proportional amounts with respect to each other. Using normative forms like "proportional" is an example commonly accepted, standard language. However, a student might also describe proportionality without using the term "Proportional" to say something like: Every time variable B changes by 1, variable A changes by 2. This is an example of a non-standard way of describing proportional relationships. If using GAI to generate possible responses for response process-based validity studies, these non-standard ways of arriving to a correct response should be reflected in GAI-generated responses.

Validity based on scoring process

We believe that a new source of validity evidence needs to be specified in the context of using GAI for assessment. This evidence source is related to response process-based validity but is focused on the GAI scoring process rather than the learner's response process. Specifically, it is becoming increasingly common to use GAI to score student assessments (Baidoo-Anu and Ansah, 2023; Moorhouse et al., 2023; Mao et al., 2024). As discussed above, historically, the validity of AI-based scores has been evaluated using human-machine agreement measures, which is related to criterion-based validity (discussed below). With the emergence of GAI models, much smaller previously labeled data sets might be needed for training the model (but this claim also needs to be further investigated empirically). It is reasonable to suggest that GAI models will need very few or no previously scored student responses to be able to score assessments that exhibit high human-GAI agreement. However, as discussed above, the agreement measures that are evidence of criterion-based validity are not suitable evidence for multiple purposes, such as those required to guide the learning process. Specifically, the agreement measures do not provide validity evidence for evaluating whether the GAI considered the same attributes in student responses to assign specific scores as a human scorer would. However, this information is necessary for ensuring that GAI models score the types of knowledge and skills that indicate knowledge application ability as a human scorer would. Otherwise, the results of GAI-scored assessments cannot be used to support students in developing knowledge application skills. Therefore, we believe it is necessary to introduce a new source of validity evidence that needs to be evaluated. We call it *GAI scoring process-based validity evidence*, which relates to assessing the alignment between human and GAI-scored response features.

This type of validity evidence parallels response process-based validity but emphasizes the need to evaluate whether the non-human scorer uses the same attributes to assign a score as a human scorer. One way to assess the *scoring process-based validity of GAI-produced* scores is to supply the GAI model with a scoring rubric focused on the relevant elements of student responses and ask the GAI to score a sample of student responses using the rubric. Next, one should ask the GAI model to explain why specific scores were assigned based on the provided rubric. This process will allow us to gauge whether the GAI model uses the same criteria for assigning scores. It is also possible to further train the GAI model and improve the scoring process-based validity through careful prompt generation and guiding the model to evaluate specific attributes of interest when scoring student responses. The steps of this method could be presented as evidence for the scoring process-based validity. In the example discussed above, we request that GPT provides a rationale for assigning LP level for each response on MSS LP-aligned assessment. Through this process, we are evaluating whether GPT is using the same rationale for assigning a score as human scorers. In cases when the rationale differs, we proceed by supplying more of the relevant examples and further revising the prompt to help GPT better align to human rationale for score assignment. This process results in improving the theoretical basis for human-GAI agreement driven by the MSS LP and therefore helps improve *scoring process-based validity* of the resulting GAI-produced scores.

In situations with no scoring rubric or LP available, one might ask GAI to develop a rubric based on criteria necessary for evaluating the relevant attributes in student responses. These attributes might be specified based on prior work on defining the construct of interest. Through careful prompt generation, GAI could be guided in developing a rubric that evaluates all the necessary attributes. This rubric can then be used to score a sample of student responses, and GAI's rationale might be asked to suggest specific scores based on the rubric. This training can serve as evidence for scoring process-based validity of the resulting GAI-based scores.

Evidence based on internal structure

Evidence based on internal structure pertains to evaluating the degree to which the relationships between items on the test relate to the construct being measured (Eignor, 2013). For example, the cognitive model that guides test development (or any conceptual framework used to design the test) might imply that the test is unidimensional. In this case, evidence should be presented that the test items conform with the theoretically suggested unidimensional structure, which will serve as evidence to suggest that the test measures the construct of interest. Alternatively (or in addition), the cognitive framework might imply that the test items measure different proficiency levels—as in the case of assessments that measure student progress along the LP levels. In these situations, evidence must be presented to show that the items on the test measure various proficiency levels in a way suggested by the LP. Examples of studies on internal latent structure validation include Kaldaras et al. (2021a,b).

If GAI models are leveraged to score assessments, evaluation of internal structure reflected in GAI-based assessment scores should also be evaluated. This could be done by applying traditional methods for assessing internal latent structure—such as latent variable modeling approaches like confirmatory and exploratory factor analysis—to GAI-produced scores to ensure that they reflect the same latent structure as human-based scores. A sample study focused on evaluating the internal latent structure of ML-generated scores was done by Kaldaras and Haudek (2022). In this study, the authors used confirmatory factor analysis to gauge the similarity between the item difficulty parameters produced using human and machine-generated scores. This approach allowed authors to identify specific items and LP levels that exhibited significant discrepancies between human and machine-assigned scores. This led to considerably different values estimated for the difficulty parameters. These results help further investigate where the AI-based scores approximate the same latent structure for a given assessment instrument, what discrepancies occur, and for which items, which is an essential aspect of the internal structure-based validation process. While this study was performed using supervised ML-based scores, similar studies can also be conducted using GAI-based scores. The CFA analysis can be easily performed using standard statistical packages such as SPSS, Lavan package for R or MPlus.

Some studies of the internal latent structure are also designed to show whether items function differently with different student populations (racial, ethnic, or gender subgroups). In this context, differential item functioning (DIF) might indicate multidimensionality that might or might not be desirable based on the framework used to guide the test development. Suppose GAI models are leveraged to score assessments. In that case, differential item functions in GAI-based assessment scores should also be evaluated to ensure that the DIF does not result from biases inherent in the GAI models.

Further, research has been done to employ machine learning approaches for identifying differential item functioning on previously designed assessments (Hoover, 2022). This study represents a promising approach for employing AI-based approaches for evaluation of DIF in various contexts using existing items and student responses. Specifically, it is important to distinguish between statistical bias (meaning a biased estimator) and the bias that reflects the influence of unintended characteristics of the examinee. Generally, if severe DIF is detected and it is not related to the target construct, the items are not used. Further studies should be conducted to refine this approach for use in practice. Analysis of DIF can be performed using SPSS, R, Mplus, Stat and SAS among others.

Another strategy for examining the latent structure of scores from both human and AI sources is employing multi-group confirmatory factor analysis (MG-CFA), which allows researchers determine whether the factor structure is different due to a scoring approach (human vs. AI) (Asparouhov and Muthén, 2014). Multigroup CFA should be done using the same set of items. The assumption is that the latent structure should stay invariant, irrespective of who (GAI or humans) scored the items. Evaluating the latent structure invariance across the two set of scores (human and AI-based) will allow to evaluate the validity of internal structure of GAI-based scores.

Evidence based on relation to other variables (test criterion and beyond)

In many situations, the intended interpretation for a given use of test scores implies that the construct should be related to other variables, which in turn requires a careful analysis of the relationship of the test scores to external variables. Since ML models have historically been used to perform the work of a human scorer, it is not surprising that the most common source of validity evidence evaluated for ML-based scores is various measures of human-machine agreement. In this context, the human scores were considered the gold standard against which the performance of the ML algorithm was evaluated. Historically, supervised ML approaches that required pre-training using previously scored data sets have been using various methods for assessing human-machine agreement (see Zhai et al., 2020 for detailed review). Previously described approaches include using the same data to train and evaluate the performance of ML algorithm (self-validation), splitting the data set into a training and testing sets (split-validation), and splitting the data set into n subsets each subset is used to train the ML algorithm while other subsets are used as a testing set to validate the model accuracy (cross-validation). Similar approaches could be used with GAI models, and likely, GAI models will require much smaller data sets (although this suggestion remains to be tested empirically).

Holding GAI to the same scoring standards as human scorers

As discussed above, supervised ML scores are usually compared to human scores, therefore establishing human scores as "the gold standard." However, this is not always the case, since it takes significant effort to ensure high-quality human scores. Unless properly trained, one should not assume that humans give you valid inferences about students. Research has shown that humans are biased toward longer than shorter responses, and therefore, human scores also represent nonperfect criteria. So, can an AI-based scoring system with nonperfect criteria be better than nonperfect criteria (human scores)? It makes sense to hold ML algorithms to the same standards as humans and evaluate these algorithms according to similar training criteria. In other words, we should replicate what is being done to train humans to replicate the high-stakes training of humans. We can refer to literature on training people to score open-ended assessments and try to replicate that process with machine algorithms. For example, seeding in previously scored responses into the scoring process to see if people are drifting away on the scoring process, then retrain them if they drift away too far—the same can be done with ML algorithms.

Further, we could build on the previously discussed split-validation method and combine it with purposeful manipulation of the training sets to study the outcomes. For example, the training sets could be selected to have responses with the same score as the human raters. In that case, the training set does not have incorrect responses (but it will have variations in student work); the training set will get perfect results.

Training AI algorithms to recognize diversity of human thinking

The amount of variation in the training set and what one chooses to vary in the training set will also affect the ML algorithm. Manipulating composition and variability in responses in the training sets can provide insight into how well a given algorithm picks up on the diversity of human thinking. This could provide evidence for the validity of AI-based scores for different types of reasoning represented in student responses. It can also help establish criteria on how much variation one needs to have in the training set to train a given algorithm to pick up consistently on this variation. For example, classifying learners into LP levels calls for having a rich distribution at all LP levels. Continuing the example with leveraging ChatGPT model for scoring student responses of MSS LP-aligned items, we discovered that having at least one representative example of student response for each type of reasoning is needed by GPT in order to assign a score is necessary for achieving high GAI-human agreement.

Another example is bias: if you have responses that are good but have a lot of spelling errors that are still given high scores, the ML algorithm might not score those properly. In fact, when using GAI for scoring MSS LP-aligned assessments, we discovered the non-standard language to be the central issue in producing mis-scores at higher levels of the LP. Specifically, we discovered that responses consistent with sophisticated reasoning but use non-standard language or show evidence of responders being non-native English speakers tended to be mis-scored by GAI to lower LP levels. A possible way of dealing with this shortcoming might be developing a vocabulary for non-standard language and using this vocabulary as part of the GAI training process. Further, training data sets can also be selected to minimize possible DIF or cultural bias or to test for sensitivity to irrelevant features of responses.

Evaluating consistency of GAI-based scores

Another way of dealing with potential misscores due to various reasons might be to ask GAI to produce scores on the same data set several times to evaluate consistency of GAI-generated scores across trials. If the GAI-based scores are consistent across the trials and agree with human-based scores, this provides evidence of criterion-based validity. The GAI-based scores that are inconsistent and disagree with human-assigned scores should be further examined to explore how the prompt might be changed to achieve better criterion-based validity. In the project on GAI scoring of MSS LP-aligned assessments, we used this approach by asking GAI to score each response three times and comparing the produced scores to human scores on the same items. In cases where we saw disagreement, we discovered that additional prompt revisions were needed to clarify the scoring approach for GPT, and better agreement was achieved as a result.

Further, one could incorporate a feature that would allow it to stop the GAI algorithm when it encounters an outlier with non-standard language or a response that is scored inconsistently across trials. All GAI-based scores should have a level of confidence, and people should be critical when interpreting the scores. The accuracy of the information is only as good as the training set. Also, one could use multiple GAI algorithms to validate and inform the validation of each other (like confirmatory and exploratory factor analysis). For example, one could use one GAI algorithm to identify patterns in a given data set (GAI 1). At the same time, one would use another GAI algorithm (GAI 2) to score the same data set and assign scores. The GAI 1 will find clusters that are considered similar by that algorithm, so they will get the same score once you score them. Compare the clusters identified by GAI 1 with scores assigned by GAI 2. Seeing how the results match up to GAI 2-ask GAI 1, what does it take to get this score? Does the response make sense? Evaluate the differences between the two algorithms to see how valid and consistent the scoring outcomes are with respect to scoring the construct of interest.

Validity generalization

An important issue in educational settings is the degree to which the validity evidence based on test-criterion relation can be generalized to new situations without further studies on validity in those new situations. This point is critical considering the push toward a wide use of GAI algorithms in educational settings for assessment purposes and beyond. When investigating the generalizability of GAI-based scores, it is essential to study to what extent GAI-based outputs generalize to situations beyond a given study or context. For example, when GAI models are used to predict scores on the same assessment items used in the original validation study, the generalizability and prediction accuracy will likely be very high. However, suppose the assessment items closely resemble those used in the validation study or are entirely different but assess a similar or the same construct. In that case, the behavior of GAI models needs to be further studied to investigate how well these models predict student performance on such assessments. Approaches such as those discussed in the previous section could be used to study the performance of GAI algorithms with new sets of student responses or with different but closely related items. Evidence gathered on the performance of these algorithms under these various circumstances could serve as evidence of generalizability.

Further, it is important to consider the drawbacks of GAI algorithms, such as hallucinations and AI drifting, in the context of generalizability studies. The problem of AI hallucinations refers to AI providing incorrect predictions that may occur even after training. AI drifting refers to situations where the accuracy of predictions produced from new input values "drifts" away from the performance during the training period. These drawbacks suggest that the outputs of GAI models should be periodically monitored and checked even after the GAI model has been released for use by the public to ensure that such drifting or incorrect predictions do not occur.

Discussion

As mentioned at the beginning of this paper, no test can ever be fully validated. Instead, a sound validity argument integrates various sources of validity evidence into a coherent account of the degree to which the available evidence and theory support the intended interpretation of the test results for specific uses. As such, a validity argument should incorporate multiple sources of validity evidence from multiple studies and show how the findings align with previously reported results if available. Validation is an iterative process that might involve revisions in the test, the associated rubrics, and the definition of the underlying latent construct. In theory, the validation process never ends as there is always additional information that can be gathered to understand the construct, the test, and the inferences that can be made more fully from the test. However, in practice, at some point, the validation process aimed to support evidence for the intended interpretation of the test results must end at least till new evidence emerges that would question the previous validity inferences in some way. The amount and type of evidence required to support specific inferences depends on many factors, including the type and goals of the test, knowledge domains, and topic advances. Higher stakes require higher evidence standards.

In the context of using GAI for various validation purposes discussed in this paper, it is essential to recognize that GAI is a continuously evolving field. This important feature of GAI algorithms has implications for the validity studies conducted with the help of GAI. For example, the GAI models are constantly learning new information, improving their overall accuracy, and increasing the range of tasks they can successfully perform. This implies that GAI algorithms can potentially identify certain instances (for example, patterns in student responses) that do not align well with the previously validated construct. In these cases, GAI can provide additional evidence requiring possible refinement of the construct definition and changes to the associated test items and rubrics. Further validation studies might be needed to support the inferences desired to be made from the test. This is just one example, and other implications might be possible because of the evolving nature of the GAI models.

In addition, very little is known about the long and even short-term effects of using GAI algorithms to solve various problems in education. This has significant consequences for validity studies conducted with the help of GAI as well. For example, considering that GAI algorithms can exhibit drifting and hallucinations (discussed above), it is essential to ensure that GAI algorithms are producing consistently accurate and reliable results in the long run. This might require constant monitoring by humans to evaluate the validity of GAI outputs for specific purposes. This is especially important if these GAI algorithms will be used to guide multiple aspects of the learning process, including aiding in assessment design and evaluation for both summative and formative purposes, and adjusting the learning process based on the results of these assessments. In each of these cases, sufficient evidence needs to be presented that GAI-based outputs produce accurate and reliable outputs and that these outputs can be used to make the desired decisions about the learning process.

Related to the previous point, it is important to consider the unintended consequences of using GAI-produced outputs to guide the educational process. In this context, the issue of bias is important to consider. As mentioned above, it is important to distinguish between statistical bias in estimation and bias in items and scoring due to influences other than the target construct. Importantly, identifying bias is always a challenge because "bias free" criterion is needed for comparison. For example, in a typical DIF study, it is assumed that most items are unbiased, so the scores from those items can be used to identify potential bias in studied items. The same would need to be true for the study of bias in GAI scoring or test development. In a sense, there needs to be a "bias free" training set so that bias can be detected when the training set is not bias free. This is especially important since GAI algorithms are being trained on large amounts of various types of human-generated data, it is important to consider the biases that could be present in the data and, therefore, become inherent in the GAI algorithms as a result. It is important to investigate the presence of these biases and their potential effects on interpreting the test results. For example, as discussed above, one should investigate potential biases of GAI algorithms based on gender background (including academic, ethnic, racial, and linguistic, among others) and their effect on GAI model outputs as well as the unintended consequences of those outputs when it relates to the interpretation of assessment results for specific purposes. For example, recent studies have shown that human and machine-based scores exhibit similar amounts of bias and suggested that diverse groups of human experts should be used to evaluate the presence of potential biases (Belzak et al., 2023). We also believe that while GAI can perform many of the tasks outlined above, the end judge of the validity of GAI actions should always be humans.

We hope that the discussion points provided in this short paper can serve as a basis for starting the conversation about establishing the standards for validity in the era of widespread use of GAI in education and educational evaluation.

Author contributions

LK: Conceptualization, Writing – original draft. HA: Conceptualization, Writing – review & editing. MR: Conceptualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The publication fee for this article was covered by the University of Colorado Boulder.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equ. Model. Multidiscip. J.* 21, 495–508. doi: 10.1080/10705511.2014.919210

Baidoo-Anu, D., and Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI* 7, 52–62. doi: 10.61969/jai.1337500

Belzak, W. C., Naismith, B., and Burstein, J. (2023). "Ensuring fairness of human-and AI-generated test items" in *International Conference on Artificial Intelligence in Education*. Springer Nature Switzerland, Cham. 701–707.

Brown, N. J., and Wilson, M. (2011). A model of cognition: the missing cornerstone of assessment. *Educ. Psychol. Rev.* 23, 221–234. doi: 10.1007/s10648-011-9161-z

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. Available at: https://hdoran.github.io/Blog/ContentMapping.pdf

Duschl, R., and Hamilton, R. (2011). "Learning science" in Handbook of Research on Learning and Instruction. Eds. R. E. Mayer and P. A. Alexander (New York: Routledge), 92–121.

Eignor, D. R. (2013). "The standards for educational and psychological testing" in APA Handbook of Testing and Assessment in Psychology, Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology. eds. K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel and S. P. Reiseet al. (Washington D.C.: American Psychological Association), 245–250.

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

Gierl, M. J., and Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Appl. Psychol. Meas.* 42, 42–57. doi: 10.1177/0146621617726788

Hoover, J. C. (2022). Using machine learning to identify causes of differential item functioning. Doctoral dissertation. University of Kansas).

Kaldaras, L., Akaeze, H., and Krajcik, J. (2021b). A methodology for determining and validating latent factor dimensionality of complex multi-factor science constructs measuring knowledge-in-use. *Educ. Assess.* 26, 241–263. doi: 10.1080/10627197.2021.1971966

Kaldaras, L., Akaeze, H., and 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. *J. Res. Sci. Teach.* 58, 589–618. doi: 10.1002/tea.21672

Kaldaras, L., Akaeze, H. O., and Krajcik, J. (2023). Developing and validating an next generation science standards-aligned construct map for chemical bonding from the energy and force perspective. *J. Res. Sci. Teach.* 1–38. doi: 10.1002/tea.21906

Kaldaras, L., and Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. *Front. Educ.* 7:968289. doi: 10.3389/feduc.2022.968289

Kaldaras, L., and Krajcik, J. (2024). "Development and validation of knowledge-in-use learning progressions" in *Handbook of Research on Science Learning Progressions*. Eds. H. Jin, D. Yan and J. Krajcik (New York: Routledge). pp. 70–87.

Kaldaras, L., and Wieman, C. (2023). Cognitive framework for blended mathematical sensemaking in science. *Int. J. STEM Educ.* 10, 1–25. doi: 10.1186/s40594-023-00409-8

Kaldaras, L., Yoshida, N. R., and Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Front. Educ.* 7:983055. doi: 10.3389/feduc.2022.983055

Krajcik, J. S. (2021). Commentary—applying machine learning in science assessment: opportunity and challenges. *J. Sci. Educ. Technol.* 30, 313–318. doi: 10.1007/s10956-021-09902-7

Kulgemeyer, C., and Schecker, H. (2014). Research on educational standards in German science education—toward a model of student competences EURASIA. *J. Math. Sci. Technol. Educ.* 10, 257–269. doi: 10.12973/eurasia.2014.1081a

Mao, J., Chen, B., and Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends* 68, 58–66. doi: 10.1007/s11528-023-00911-4

Messick, S. (1980). Test validity and the ethics of assessment. *Am. Psychol.* 35, 1012–1027. doi: 10.1037/0003-066X.35.11.1012

Ministry of Education, P. R. China (2018). Curriculum Plan for Senior High School. Beijing: People's Education Press.

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003, i–29. doi: 10.1002/j.2333-8504.2003.tb01908.x

Moorhouse, B. L., Yeo, M. A., and Wan, Y. (2023). Generative AI tools and assessment: guidelines of the world's top-ranking universities. *Comput. Educ. Open* 5:100151. doi: 10.1016/j.caeo.2023.100151

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

National Research Council (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academies Press.

NGSS Lead States (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.

OECD (2016). PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy. Paris: OECD Publishing.

Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academy Press.

Samala, A. D., Zhai, X., Aoki, K., Bojic, L., and Zikic, S. (2024). An in-depth review of ChatGPT's pros and cons for learning and teaching in education. *Int. J. Interact. Mob. Technol.* 18, 96–117. doi: 10.3991/ijim.v18i02.46509

Yao, J. X., and Guo, Y. Y. (2018). Core competences and scientific literacy: the recent reform of the school science curriculum in China. *Int. J. Sci. Educ.* 40, 1913–1933. doi: 10.1080/09500693.2018.1514544

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., and Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Stud. Sci. Educ.* 56, 111–151. doi: 10.1080/03057267.2020.1735757