

Same Score, Same Meaning? Generative AI, Automated Scoring, and the Validity of Score Interpretations

Christopher Runyon

NBME

crunyon@nbme.org

Abstract

A necessary, but not sufficient, condition for using an automated scoring system is that automated scores strongly correlate with subject-matter expert scores. One must also demonstrate that the automated scoring system derives its score based on construct-relevant features. However, for general-purpose foundational generative AI (GAI) models, one often does not have access to the training data or procedures used to build the model. Drawing on Kane's scoring inference, I argue that this GAI model development opaqueness is a validity threat, and under some conditions limits the attribution of subject-matter expert score meaning to GAI automated scores, even when the scores agree. I distinguish text-proximal from expertise-dependent scoring tasks, walk through an example of scoring a common performance assessment in medical education with analytic and holistic rubrics, and identify the resulting implications for interpretations and use.

1 Introduction

The use of generative artificial intelligence (GAI) has permeated many aspects of medical education (Gordon et al., 2024; Nagi et al., 2023), with the range of use cases expanding more quickly than corresponding education about how GAI models work. Several helpful articles provide guidance on what users should do to effectively use GAI models (Tolsgaard et al., 2023; Kim et al., 2025; Nagi et al., 2023); however, much of this guidance lacks more technical explanations of GAI architecture, including the functional mechanisms by which these models are trained and generate responses. How much this knowledge matters depends on the use case. For low-stakes uses, such as improving the clarity of a draft email, these details may be largely irrelevant. For generating or scoring assessment content or aiding high-stakes decisions, knowledge

of how GAI systems function becomes important for judging whether GAI use is appropriate.

This paper examines how GAI model training and architecture affect a validity argument when these models are used to score performance assessments in an automated scoring system. The discussion specifically focuses on aspects of Kane's (1992, 2013) scoring inference through a sample use case of scoring objective structured clinical encounter (OSCE) notes with analytic and holistic rubrics (Sloan et al., 1995; Yune et al., 2018; Jönsson and Balan, 2018). Central to this discussion is the relationship between construct representation and the text proximity of the performance being scored. I argue that for tasks where the construct representation is more text-proximal, such as scoring assessments with highly explicit analytic rubrics, there may be strong alignment between the capabilities of GAI models and intended score meaning. In contrast, for tasks where the construct representation is less text-proximal and requires deep subject-matter expertise, such as scoring with a holistic rubric that requires contextual clinical judgment, GAI models may lack the appropriate grounds for supporting the intended score meaning.

An appropriate step in developing any automated scoring system is to ensure that the score being produced by the automated scoring system correlates with subject-matter expert (SME) scores when scoring the same assessment (Attali and Burstein, 2006; Bejar, 2011; Williamson et al., 2012). While necessary, this score correlation is not a sufficient condition for inferring that the scores generated by the automated scoring system and the SME scores are equivalent in meaning. If sufficient information about the automated scoring system's training data and processes are not available (such as with many proprietary GAI models), then the automated scoring system may lack the appropriate epistemic grounds to produce a score that represents an expert's judgment. I challenge the idea

that just because SMEs and GAI-powered automated scoring systems produce similar scores, that the corresponding score meanings are equivalent.

I call this the Fallacy of Algorithmic Knowledge Equivalence in Score MEaning, or FAKE SME: the mistaken inference that because a GAI system can generate scores resembling SME judgments, the meaning of those scores is also transmitted through the generation. This inferential fallacy has direct implications for score use and consequences, especially when scores are used to support judgments about high-stakes assessments or decisions.

The manuscript proceeds as follows. I first review the validity concepts that motivate the argument, focusing on Kane's (1992, 2013) scoring inference and the roles of warrants and backing, and on Williamson et al. (2012)'s emphasis on the conceptual fit between construct and scoring features. I then describe how GAI-based scoring restructures the scoring inference, discuss OSCE-note scoring with analytic and holistic rubrics, and conclude with implications for interpretations and use.

1.1 Kane's Scoring Inference and Automated Scoring

Kane's argument-based validity framework has been extremely influential in educational measurement (Kane, 1992, 2013). This theory utilizes an interpretive chain of sequential inferences, including scoring, generalization, extrapolation, and use. Each inference entails its own warrant, meaning the assumptions that must hold for the inference to obtain, and backing, meaning the empirical evidence and theoretical support for those warrants. Although each inference is important, the discussion focuses primarily on the scoring inference because this is the point at which observed performance is translated into an observed score.

The scoring inference answers the question: Does the assigned score adequately represent the quality of the performance according to the scoring rules? When experts are used to score a performance assessment, warrants for the scoring inference include that the SME has the requisite expertise, can apply the rubric appropriately, and bases the score on construct-relevant features. Corresponding backing includes evidence of rater training, calibration, appropriate interrater reliability metrics, and validity evidence for the construction of the scoring rubric. In this setting, the SME acts as an instrument within the scoring inference.

When automated scoring systems are used to

score a performance assessment, the warrants and backing shift. Generally speaking, warrants include that the automated scores approximate SME scores with sufficient accuracy, that the automated score represents the same scoring construct as the SME score, and that the automated system is robust to construct-irrelevant manipulation (Attali and Burstein, 2006; Bejar, 2011; Williamson et al., 2012; McCaffrey et al., 2022; Ferrara and Qunbar, 2022). Corresponding backing includes evidence of model training procedures, the selection and quality of training data, expert-system scoring agreement statistics, robustness testing, and the conceptual fit between the scoring model and the construct.

An additional distinction concerns the kind of epistemic backing available for each scoring process. When SMEs score a performance assessment, it is typically possible to identify how their scoring expertise was acquired. For example, in medical education, a physician-educator has passed through some combination of formal education, supervised practice, calibration against standards, and experience with the consequences of decisions. This history does not make SME judgment infallible, but it provides an epistemic basis for treating the SME as a plausible instrument for applying the scoring rules. In custom automated scoring systems, the epistemic backing is different but still often inspectable: developers can describe the training sample, the chosen scoring features, the modeling approach, and how these decisions relate to the intended construct (Attali and Burstein, 2006; McCaffrey et al., 2022). With GAI systems, the source of scoring capability is harder to identify. The model may be trained on text that contains relevant domain knowledge, but that knowledge is distributed across large-scale pretraining, tuning, and configuration. The GAI model is not trained for a specific scoring task and has the more general design purpose of generating text, which some have argued prevents GAI models from having the appropriate epistemic backing to make judgments *at all* (Quattrociochi et al., 2025). Thus, compared with SME and custom automated scoring systems, GAI scoring provides weaker epistemic access to how the relevant scoring expertise was acquired and utilized in the score generation.

This difference in epistemic backing extends earlier concerns in the automated scoring literature. Williamson et al. (2012)'s framework for the evaluation and use of automated scoring identifies "con-

struct evaluation” as an important component, including “a summary judgment of the conceptual fit between the scoring features of the automated system and the construct of interest” (Williamson et al., 2012, p. 6). The introduction of GAI models warrants further reflection on this point, because what has changed is the nature of the scoring system being evaluated. GAI models are qualitatively different from many of the custom automated scoring systems that preceded them. In a custom system, the training data was purposefully curated and scoring features may be intentionally selected, engineered, and evaluated in relation to the construct (Attali and Burstein, 2006; McCaffrey et al., 2022). In a GAI system, the score is generated through a general-purpose language model whose capabilities were shaped by massive, heterogeneous training data and the relevant text-generation features are designed for general-purpose use. The relevant validity question shifts to whether the GAI scoring protocol provides sufficient evidence that the generated scores preserve the intended score meaning.

1.2 What Changes When the Automated Scorer Is GAI?

The warrants and backing are different when GAI is used in an automated scoring system. In addition to producing scores that are similar to SME scores, the warrant is that the GAI system can correctly interpret the rubric within the scoring instructions and apply it to the assessment outcome in a way that resembles expert judgment for construct-relevant reasons. The GAI system must base its scores on the intended construct rather than on surface-level linguistic features, indirect cues, or construct-irrelevant patterns embedded in the rubric, task, or response, much like with other automated scoring systems (Williamson et al., 2012; McCaffrey et al., 2022). Research has already shown that AI-generated essays that are well constructed but lack strong analytical reasoning have received inflated scores with automated scoring systems (Zhong et al., 2026), highlighting this vulnerability.

A key difference between GAI and both SME scoring and custom automated scoring systems is the role of the full scoring protocol. When SMEs are applying the rubric, it is being directly interpreted in the context of the area of their expertise. In a custom automated scoring system, the rubric is developed through the integration of specific construct-relevant metrics. In a GAI sys-

tem, the rubric is consumed and represented by a prompt-mediated scoring process that combines the model’s learned language representations, the supplied rubric text, any exemplars or anchors (if provided), and the system’s generation and post-processing rules.

This distinction creates a central validity concern. GAI systems may produce a plausible score (and a plausible explanation if prompted) without actually using the rubric in the same way that SMEs do. GAI systems may still produce meaningful scores. However, the scoring inference requires different support because these systems do not apply a rubric in the same way as an SME or a purpose-built scoring model. For SME raters, the validity argument can appeal to common ways in which expertise is developed and calibrated. For custom automated scoring systems, the validity argument can appeal to the design of the scoring features, training labels, and model development process. For GAI systems, the grounding for how the model was trained to generate the score is less transparent. Even if the relevant expertise was transmitted through training, it is hard to determine that this expertise was acquired in the form or depth required for the intended scoring task.

In the example of using a GAI model to score a medical student assessment, some of the requisite professional expertise may be encoded in the GAI model’s training. Medical reasoning, clinical guidelines, scoring rubrics, case discussions, and expert commentaries all exist in textual form, so GAI models may acquire representations that support expert-like judgments to some capacity. However, this possibility does not permit the scoring inference. For most general-purpose GAI systems, users cannot directly verify whether the relevant expertise was encoded, whether it was encoded adequately, or whether it is being invoked appropriately during scoring. The epistemic backing for GAI scoring is thus weaker than it may initially appear: the model’s fluency and use of expert language provide the appearance that it has the requisite expertise, but not necessarily that it has acquired or applied the judgmental competence that gives expert scores their intended meaning (a structural phenomenon that has been called *Epistemia*; see Quattrocchi et al. 2025). Even when a model produces a score that matches an SME’s score, the agreement is underdetermined—the human-model score agreement is compatible with multiple explanations, and only some of these explanations support the intended

scoring inference.

In Kane’s framework, this uncertainty functions as a potential rebuttal to the scoring warrant when a GAI system is used for automated scoring. The proposed warrant may be that the GAI system can apply relevant professional standards to evaluate the performance, and the backing may include evidence that the system produces scores that align with expert ratings. The rebuttal is that it is unclear whether the system has acquired the relevant domain expertise or is instead approximating SME judgments through surface-level or indirect textual cues. The qualifier, then, is that confidence in the scoring inference should depend on the stakes of the decision, the explicitness of the construct as represented in the text of the rubric and assessment output, and the consequences of erroneous scores.

1.3 Text-Proximal and Expertise-Dependent Scoring Tasks

I distinguish between text-proximal and expertise-dependent scoring tasks to further motivate the discussion. Text-proximal scoring tasks are those in which much of the construct-relevant evidence is expressed in the observable text itself (a low inferential distance, in the language of Messick 1989 and Kane 2013). An example is middle-school writing fluency, which includes aspects like basic composition quality, grammar and mechanics, organization, clarity, and response completeness. For these tasks, a GAI scoring system may perform well because the relevant features are linguistically encoded, commonly described in training data, and often captured by rubrics that can be clearly communicated in natural language. The level of expertise required to accomplish these tasks may reasonably be inferred to have been transmitted through the training data and processes. Score agreement alone is again insufficient, but the underlying claim is more plausible because the scoring task is closely aligned with what language models are designed to process: text.

By contrast, expertise-dependent scoring tasks are those where the response artifact is only one piece of evidence and must be interpreted through specialized professional judgment. In medical education, examples may include scoring clinical reasoning performances, evaluating entrustable professional activities, or determining readiness for unsupervised practice. For these tasks, the construct-relevant features may be tacit, context-dependent, locally defined, or tied to standards of

practice—visible only to someone who understands the broader system in which the performance occurs. The warrant must address whether the model has access to, represents, and can appropriately apply the domain-specific standards and judgmental competence that human experts develop through traditional pathways. It takes more faith on the part of the user to infer that this level of expertise has been inherited by the GAI model. The less explicit and more expertise-dependent the judgment, the stronger the validity threat for GAI scoring, and the less sufficient numeric agreement becomes.

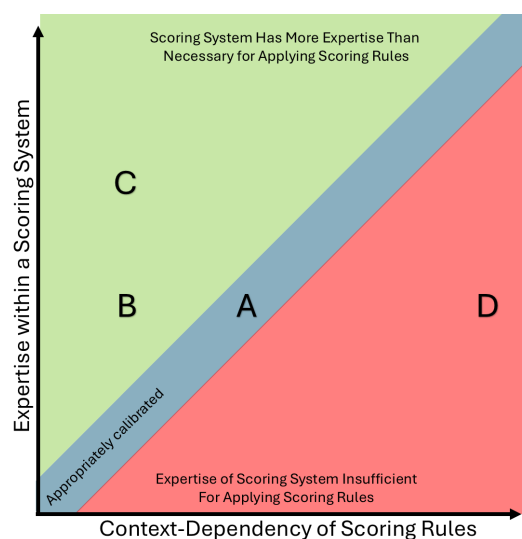


Figure 1: Relationship between scoring system expertise and scoring rule context-dependency, with illustrative examples at varying levels of calibration.

Figure 1 relates the expertise available within a scoring system to the context-dependency of the scoring rules. The diagonal represents combinations in which the two are appropriately calibrated. Point A represents a customized automated scoring system where the training data and model have been engineered to be calibrated for the level of expertise to score the assessment. Above the diagonal, a system brings more expertise than the rules strictly require; these combinations are resource-inefficient but can still function as intended and support strong validity arguments. Point B might represent the use of a GAI model to score a simple task where the model training can reasonably be inferred to be more than sufficient; Point C represents the use of SMEs to score an elementary task where their expertise far exceeds what is necessary. The combinations of concern in this manuscript fall below the diagonal, where there is insufficient validity evidence to support claims that the sys-

tem can apply the scoring rules. Point D is one such system, such as where a GAI system is being used to score an assessment that is highly context-dependent and the information necessary to score the assessment is less text-proximal. The farther a system strays below the diagonal line, the stronger the validity threat.

1.4 OSCE Note Scoring

The example of OSCE-note scoring in medical education is useful here because the same student performance can be scored for different purposes using different types of rubrics that place different demands on the scorer (Sloan et al., 1995). OSCE examinations are simulated clinical encounters with a standardized patient (a person trained to portray a specific illness) that require students to integrate medical knowledge and clinical practice, demonstrating their clinical reasoning. Often these assessments are evaluated through several lenses, including feedback from the standardized patient. A typical target of the assessment is the post-encounter note, which is a summary of the relevant information learned about the patient, along with a diagnosis and initial steps in treatment.

An analytic rubric may ask whether the student included a relevant history element, documented a key physical finding, identified an appropriate diagnosis, or included a reasonable management step. In this setting, much of the cognitive work occurs upstream, when the OSCE case designer identifies the evidence that should be present in the note if the learner engaged in the intended clinical reasoning. The scoring task then becomes, at least in part, determining whether that evidence is present in the student's written response, including lexical variants of these elements (i.e., using different words to report the same information), a task that GAI systems are well designed to accomplish.

For example, an analytic rubric criterion may be that a student note includes "two-week history of progressively worsening cough." For a GAI model prompted to scan an OSCE note to see whether this information is present, the contextual depth of an experienced physician is not required (evidence suggests that nonclinicians can effectively apply analytic rubrics to clinical assessments; Berger et al., 2012; Yudkowsky et al., 2019). The GAI model processes the instructions, the analytic rubric, and the student note as language, creating context-sensitive representations of what the student wrote and what the criterion is asking for. The model

then predicts a string of tokens that reflects the relationship between the student response, the rubric criterion, and learned patterns of how such judgments are expressed. For example, it might identify "the cough has been getting worse over the last 10 days" as a match but "the patient has been coughing for a while" as insufficiently specific.

Now consider a holistic rubric for the same OSCE note, which asks the scorer to judge whether the note is coherent and reflects appropriate prioritization, or whether the learner's performance suggests readiness for a particular level of supervision. Here, the scorer is doing more than detecting whether specific text is present in the note. The scorer must interpret the performance in context, which requires the requisite level of contextual expertise.

When clinicians apply a holistic rubric to an OSCE note, they are drawing on knowledge developed through education, clinical experience, feedback, and other traditional clinical training. This developmental history functions as part of the epistemic backing for trusting the expert's judgment. The expert connects the words in the note to a broader network of pathophysiological pathways, patient encounters, standards of care, and developmental expectations for learners. They notice not only what is present but what is missing, distinguish a minor omission from a dangerous one, recognize when a note appears organized but does not follow standards of practice, and gauge how much uncertainty is tolerable for a learner at a given level of training.

Case-specific information could be added to the holistic rubric prompt to supply context, but the validity challenge remains: the scoring inference still assumes that this additional information, combined with whatever expertise the model has inherited through training, is sufficient for the requisite expert judgment. To make such scoring defensible, concepts like "reflects appropriate prioritization" and "readiness for supervision" would need to be operationalized into prompt language that leads the GAI system's language representations to approximate the requisite expertise. Even then, the validity argument must assume that the GAI system has inherited the contextual expertise to identify relevant information omitted from a learner's performance. Such boundaries around relevant omissions are especially challenging for GAI-powered scoring inferences when contextual expertise is required.

This is where FAKE SME becomes most con-

sequential. A GAI model may generate the same holistic score as an SME. But the same numeric score does not necessarily carry the same meaning if the score was produced through a different process. For the SME, a score of 4 may mean that the performance reflects a clinically acceptable but incomplete demonstration of reasoning for a learner at a particular level. For the GAI system, a score of 4 may reflect learned associations among common medical assessment phrases or irrelevant matched patterns of language from medical texts. These processes may often converge on the same score, but the convergence does not by itself establish equivalence in score meaning.

The validity issue becomes clearer when evaluating differences between SME and GAI scores. Human–human interrater reliability is often reported alongside human–AI agreement, with similar or better human–AI agreement taken as favorable evidence that the automated system is working appropriately. But what if the GAI system is systematically producing scores that correlate with features of the response that themselves correlate with the required expertise—representing the rubric and applying scores in a way that is not based on the construct of interest, yet close enough to appear so? At least with human–human agreement, the source of a disagreement can often be identified, and in some cases the disagreement can be shown to be justified. With an opaque GAI system, all we have is faith that the relevant expertise was transmitted through training.

GAI models' internal representations are likely more sophisticated than their characterization in this argument: GAI models process complex language, represent semantic relationships, and generate outputs that appear sensitive to context. But the validity concern remains. Users still need evidence that the scoring protocol elicits representations at the depth the scoring task requires, not merely the appearance of contextual judgment.

1.5 Implications for Validation and Use

In Kane's validity framework, any validity threat present in the scoring inference propagates downstream to the subsequent inferences of generalization, extrapolation, and use; weaknesses in one inference do not remain isolated. This propagation is particularly important for GAI-powered scoring systems in highly contextualized fields such as medical education. A GAI system may produce scores that resemble expert judgments with apparently suf-

ficient numerical agreement, while still failing to apply the clinical, developmental, or contextual expertise that gives those scores their intended meaning. If this occurs, the problem may not be immediately visible in the scoring data. The score may correlate with expert ratings, reliability indices may appear acceptable, and model-generated rationales may sound clinically plausible. The consequences of a weak scoring inference may instead emerge later, when learners enter clerkships, residency, or unsupervised practice.

This creates a difficult validation problem. In principle, one might compare students scored by experts with students scored by an independent GAI-powered system and evaluate downstream outcomes years later. In practice, such a study is likely infeasible. Even if researchers could link GAI-scored OSCE notes to clerkship performance, residency evaluations, or patient-care outcomes years later, too many intervening variables would prevent isolating the GAI scoring process as a contributing factor.

Counterintuitively, I argue that this inability to definitively observe downstream effects is precisely why the backing for the scoring inference is so important. Because it is often impossible to wait for, isolate, or causally attribute downstream consequences, the validity argument must be strengthened at the point where the score is produced. The more difficult it is to observe downstream consequences directly, the more important it becomes to establish that the scoring process is construct-relevant at the outset.

Although this paper has used OSCE-note scoring as a running example, the issue of inherited expertise extends well beyond this specific use case. The same concern applies whenever a GAI system is used in place of an SME whose judgment is valued because of domain-specific or contextual expertise. This caution is therefore not limited to OSCE notes or even to assessment. When GAI substitutes for human judgment, score agreement is not enough. Unless we can show that the system's output reflects the expertise the task requires, we risk mistaking the appearance of expertise for expertise itself. We risk a FAKE SME.

References

- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

- Isaac I. Bejar. 2011. [A validity-based approach to quality control and assurance of automated scoring](#). *Assessment in Education: Principles, Policy & Practice*, 18(3):319–341.
- Alexandra J. Berger, Colleen C. Gillespie, Linda R. Tewksbury, Ivey M. Overstreet, Ming C. Tsai, Adina L. Kalet, and Jennifer B. Ogilvie. 2012. [Assessment of medical student clinical reasoning by “lay” vs physician raters: Inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination](#). *The American Journal of Surgery*, 203(1):81–86.
- Steve Ferrara and Saed Qunbar. 2022. [Validity arguments for AI-based automated scores: Essay scoring as an illustration](#). *Journal of Educational Measurement*, 59(3):288–313.
- Morris Gordon, Michelle Daniel, Aderonke Ajiboye, Hussein Uraiby, Nicole Y. Xu, Rangana Bartlett, Janice Hanson, Mary Haas, Maxwell Spadafore, Ciaran Grafton-Clarke, Rayhan Yousef Gasiea, Colin Michie, Janet Corral, Brian Kwan, Diana Dolmans, and Satid Thammasitboon. 2024. [A scoping review of artificial intelligence in medical education: BEME guide no. 84](#). *Medical Teacher*, 46(4):446–470.
- Anders Jönsson and Andreia Balan. 2018. Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research & Evaluation*, 23(12):1–11.
- Michael T. Kane. 1992. [An argument-based approach to validity](#). *Psychological Bulletin*, 112(3):527–535.
- Michael T. Kane. 2013. [Validating the interpretations and uses of test scores](#). *Journal of Educational Measurement*, 50(1):1–73.
- Do-Hwan Kim, Ye Ji Kang, and Young-Mee Lee. 2025. [Twelve tips for developing and implementing AI curriculum for undergraduate medical education](#). *Medical Education Online*, 30(1):2585637.
- Daniel F. McCaffrey, Jodi M. Casabianca, Kathryn L. Ricker-Pedley, René R. Lawless, and Cathy Wendler. 2022. [Best practices for constructed-response scoring](#). *ETS Research Report Series*, 2022(1):1–58.
- Samuel J. Messick. 1989. Validity. In Robert L. Linn, editor, *Educational Measurement*, 3 edition, pages 13–103. American Council on Education and Macmillan Publishing Company, New York.
- Fatima Nagi, Rawan Salih, Mahmood Alzubaidi, Hurmat Shah, Tanvir Alam, Zubair Shah, and Mowafa S. Househ. 2023. [Applications of artificial intelligence \(AI\) in medical education: A scoping review](#). In *Healthcare Transformation with Informatics and Artificial Intelligence*, volume 305 of *Studies in Health Technology and Informatics*, pages 648–651. IOS Press.
- Walter Quattrociocchi, Valerio Capraro, and Matjaž Perc. 2025. [Epistemological fault lines between human and artificial intelligence](#). *Preprint*, arXiv:2512.19466.
- David A. Sloan, Michael B. Donnelly, Richard W. Schwartz, and William E. Strodel. 1995. [The objective structured clinical examination: The new gold standard for evaluating postgraduate clinical performance](#). *Annals of Surgery*, 222(6):735–742.
- Martin G. Tolsgaard, Martin V. Pusic, Stefanie S. Sebok-Syer, Brian Gin, Morten Bo Svendsen, Mark D. Syer, Ryan Brydges, Monica M. Cuddy, and Christy K. Boscardin. 2023. [The fundamentals of artificial intelligence in medical education research: AMEE guide no. 156](#). *Medical Teacher*, 45(6):565–573.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.
- Rachel Yudkowsky, Abbas Hyderi, Jaime Holden, Robert Kiser, Richard Stringham, Ananya Gangopadhyaya, Asra Khan, and Yoon Soo Park. 2019. [Can nonclinician raters be trained to assess clinical reasoning in postencounter patient notes?](#) *Academic Medicine*, 94(Supplement 1):S21–S27.
- So Jung Yune, Sang Yeoup Lee, Sun Ju Im, Bee Sung Kam, and Sun Yong Baek. 2018. [Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students](#). *BMC Medical Education*, 18(1):124.
- Yang Zhong, Jiangang Hao, Michael Fauss, Chen Li, and Yuan Wang. 2026. [AI-generated essays: Characteristics and implications on automated scoring and academic integrity](#). *Educational Measurement: Issues and Practice*, 45(1):e70013.