

A Validity Threat Framework for Measuring Student Generative AI Use

Olukayode E. Apata¹, Yetunde O. Fawehinmi¹, Daniel O. Oyeniran², Glory O. Saidu³, Segun T. Ajose⁴, Naphtali Onalo⁵, Oluwasegun M. Amoniyan⁶

¹College of Education and Human Development, Texas A&M University

²College of Education, University of Alabama

³College of Education and Human Development, University of Minnesota

⁴College of Education, University of Kentucky

⁵Faculty of Education, University of Ibadan

⁶Kenneth P. Dietrich School of Arts & Sciences, University of Pittsburgh

Corresponding author: apata.olukayode@gmail.com

Abstract

This conceptual paper presents a validity threat framework for measuring students' GenAI use in higher education. It identifies threats related to construct definition, response processes, self-report bias, policy context, fairness, and score interpretation, offering practical guidance for stronger measurement, assessment, policy, and pedagogy in AI-integrated education.

1 Introduction

The growing use of generative artificial intelligence (GenAI) in higher education has created new questions about student learning, academic integrity, assessment design, and institutional policy (Apata, Kwok, & Ajose, 2026; Malik et al., 2024). Students may use GenAI tools to brainstorm ideas, revise writing, summarize readings, generate explanations, support coding tasks, prepare for examinations, or complete course assignments (Li et al., 2025). Although these practices are now central to debates about teaching and assessment, the measurement of student GenAI use remains conceptually unstable (Kofinas et al., 2025). A single question about whether students “use ChatGPT” can combine different behaviors, purposes, degrees of reliance, and ethical meanings (Qu et al., 2025).

This measurement problem matters because data on student GenAI use are often linked to consequential interpretations about learning, academic integrity, assessment validity, and institutional policy (Gruenhagen et al., 2024; Miao & Holmes, 2023). Survey findings may be used to describe patterns of adoption, evaluate academic integrity risks, guide instructor decisions, justify policy responses, or design AI literacy

interventions (Miao & Holmes, 2023). However, if the measurement of student AI use is vague, the resulting scores may not support the claims being made from them (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1989). For example, frequent GenAI use may reflect productive study support, routine writing assistance, limited experimentation, professional skill development, or inappropriate outsourcing of academic work (Gruenhagen et al., 2024; Lund et al., 2025). Without careful attention to validity, these distinct practices may be treated as the same phenomenon (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1989).

A validity-centered approach is therefore needed. Contemporary validity theory emphasizes that validity is not simply a property of a test, scale, or survey instrument. Rather, validity concerns the degree to which evidence and theory support the interpretations and uses of scores (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1989). Kane's argument-based approach further emphasizes that validation involves evaluating the plausibility of claims made from scores and the assumptions that connect observed responses to intended interpretations (Kane, 2013). Applied to GenAI use, this means that researchers must ask what the measure is intended to represent, how students understand the questions, what contextual factors shape their responses, and what conclusions can be drawn from the resulting data.

This paper contributes to AI measurement scholarship by shifting the discussion of student GenAI use from a general behavioral-adoption question to a validity framework that researchers, instructors, and institutions can use to design

clearer measures, interpret student AI-use data more cautiously, and avoid unsupported claims about learning, integrity, or policy compliance. The framework identifies five domains of threat: construct definition, response process, self-report, contextual and policy, and interpretation and use. The paper then translates the framework into practical measurement considerations for assessment, policy, and pedagogy in AI-integrated higher education.

2 Conceptual Background

GenAI tools have expanded the range of ways students can engage with academic tasks (Yan et al., 2024). Unlike earlier digital tools that primarily supported information retrieval, word processing, or communication, GenAI systems can produce explanations, examples, outlines, summaries, translations, code, feedback, and extended written responses (Giannakos et al., 2025). UNESCO's guidance on GenAI in education and research emphasizes policy planning, regulation, and human capacity development so that GenAI supports a human-centered vision of education (Miao & Holmes, 2023), which aligns with peer-reviewed arguments for human-centered design and oversight in GenAI-supported learning (Yan et al., 2024).

One difficulty is that “use” can refer to many different actions. A student who asks a GenAI tool to explain a statistical concept is engaging in a different behavior from a student who submits AI-generated text with little personal contribution. Similarly, a student who uses AI to correct grammar is not necessarily using it in the same way as a student who uses AI to solve an entire problem set. When these behaviors are measured with broad survey items, the resulting data may obscure important distinctions among learning support, productivity support, professional preparation, and academic misconduct (Miao & Holmes, 2023; Gruenhagen et al., 2024).

A second difficulty concerns the ethical meaning of GenAI use. In many higher education settings, students receive mixed messages about whether AI use is permitted, encouraged, discouraged, or prohibited (Gruenhagen et al., 2024; Lund et al., 2025). Some instructors may allow AI for brainstorming but not for writing. Others may allow editing support but not content generation. In some courses, AI use may be considered part of developing professional skills,

while in others it may be treated as a threat to academic integrity. These differences mean that the same behavior can have different meanings depending on the course, task, discipline, and policy context.

A third difficulty is that common indicators, especially frequency of use, are often insufficient. Knowing how often a student uses GenAI does not necessarily reveal why the student uses it, how deeply the student relies on it, whether the use is permitted, whether it supports learning, or whether it violates academic expectations. Frequency may be useful, but it should not be overinterpreted. Measurement of student GenAI use requires attention to type of use, purpose of use, task context, policy context, access, and the claims that researchers intend to make from the data (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1989).

3 Conceptual Approach

This paper uses a conceptual synthesis to develop a framework for identifying validity threats in measures of student GenAI use. The synthesis draws from contemporary validity theory, especially the Standards for Educational and Psychological Testing, Messick's unified validity framework, and Kane's argument-based approach to validation (American Educational Research Association et al., 2014; Kane, 2013; Messick, 1989). It also draws from research on response processes, cognitive interviewing, sensitive survey questions, common method bias, digital inequality, and recent scholarship on GenAI in higher education. The framework was developed by asking where claims about student GenAI use may break down when researchers move from observed survey responses to broader interpretations about learning, academic integrity, assessment quality, or policy compliance. Five recurring points of vulnerability were identified: whether the construct has been defined clearly, whether students understand items as intended, whether self-reports accurately reflect behavior, whether contextual and policy differences affect score meaning, and whether researchers or institutions use scores to support claims that exceed the evidence. These points form the five validity threats presented below. In this sense, the framework functions as a preliminary construct map for student GenAI use by identifying the dimensions that should be made explicit before scores are interpreted: tool or system, academic task, purpose of use, degree of reliance, policy status, response process, and intended score use. Because this is a

conceptual paper, the framework should not be read as a completed validation model or a fixed checklist. Instead, it offers a structure for designing, reviewing, and interpreting measures of student GenAI use. Its purpose is to help researchers make their validity argument explicit by connecting construct definitions, item design, evidence of response processes, contextual information, and intended score use.

4 Five Validity Threats in Measuring Student GenAI Use

The central conceptual claim of this paper is that measures of student GenAI use are vulnerable to five related validity threats. These threats are not isolated problems. They interact across the measurement process, from construct definition to score interpretation. Together, they show why broad or frequency-based indicators alone are not sufficient for defensible claims about student GenAI use.

4.1 Construct Definition Threat

The first threat concerns the definition of the construct itself. In many studies, “student use of GenAI” may be treated as a single, clear behavior. However, GenAI can support many academic functions, including personalized learning support, feedback, assessment support, learning design, learning regulation, automated content generation, brainstorming, summarizing, editing, translating, explaining difficult concepts, generating practice questions, writing code, creating outlines, revising paragraphs, and producing complete assignment responses (Yan et al., 2024; Giannakos et al., 2025). These functions differ in purpose, level of student involvement, ethical meaning, and relationship to learning because GenAI can operate either as a learning aid or as a substitute for students’ own cognitive work (Yan et al., 2024; Kizilcec et al., 2024).

When these different practices are combined under a broad measure of “AI use,” the resulting score may lack conceptual clarity. A student who uses AI as a study tutor may receive the same score as a student who uses AI to complete an assignment with little personal effort. In this case, the measure may not distinguish learning support from academic outsourcing. This threatens validity because the score may represent a mixture of behaviors rather than a clearly defined construct (Kane, 2013; Sireci & Faulkner-Bond, 2014).

A construct-definition threat also occurs when researchers fail to distinguish among related but non-identical constructs, such as access,

awareness, frequency, reliance, purpose, ethical judgment, and quality of use. Treating these dimensions as a single form of AI engagement may weaken score interpretation, as the resulting scores may not clearly represent the construct researchers intend to measure (American Educational Research Association et al., 2014; Kane, 2013).

For measurement purposes, student GenAI use should therefore be treated as a multidimensional construct rather than a single behavior. Measures should clarify whether they focus on access, awareness, frequency, reliance, learning support, writing support, content generation, assessment completion, feedback seeking, disclosure, or policy compliance. Researchers do not need to measure all forms of use in every study. However, they do need to state which dimension of GenAI use is being measured, why that dimension matters, and what interpretation the resulting score can support (American Educational Research Association et al., 2014; Kane, 2013).

4.2 Response Process Threat

The second threat concerns how students understand and respond to measurement items. Even when researchers define GenAI use carefully, students may interpret survey questions differently from what the researcher intends. This is especially likely when items include terms such as “AI use,” “misuse,” “ethical use,” “academic integrity,” “assistance,” or “unauthorized support.” These terms may appear straightforward to researchers, but students may attach different meanings to them depending on their experiences, courses, discipline, and institutional policies. This is because survey responses depend on how respondents understand, retrieve, judge, and report information; unclear GenAI items may produce responses that reflect differences in interpretation rather than actual differences in behavior (Tourangeau et al., 2000).

For example, an item asking whether a student has used AI for academic work may be interpreted narrowly by one student as using ChatGPT to generate written responses. Another student may interpret the same item broadly to include Grammarly, Google Translate, Microsoft Copilot, AI search features, citation tools, or automated feedback systems. A third student may not know whether embedded AI features in common software count as GenAI. In this situation, differences in responses may reflect differences in

item interpretation rather than true differences in behavior (Padilla & Benítez, 2014).

For GenAI measurement, evidence of the response process is especially important because students may not share a common understanding of what counts as AI assistance. Clear definitions, examples, cognitive interviews, pilot testing, and open-ended follow-up questions can strengthen evidence that students understand the measure as intended (Beatty & Willis, 2007; Padilla & Benítez, 2014; Willis, 2005).

4.3 Self-Report Threat

The third threat concerns the accuracy of self-reported data. Much of the research on student GenAI use relies on surveys, but self-reports are vulnerable to underreporting, overreporting, recall error, and common method bias. This concern is especially serious because GenAI use is often connected to academic integrity, assessment assistance, and perceptions of misconduct (Gruenhagen et al., 2024). Students may worry that admitting certain forms of AI use could lead to judgment, punishment, or suspicion. As a result, they may minimize or deny their use, even when the study is anonymous.

Underreporting may occur when students believe that AI use is prohibited, morally questionable, or likely to be interpreted as cheating. Overreporting is also possible. Some students may exaggerate their GenAI use to appear technologically skilled, innovative, or aligned with current digital trends. Others may overreport because they count many forms of digital assistance such as AI, including tools that the researcher may not intend to include. These concerns are consistent with research on sensitive survey questions and social desirability bias (Krumpal, 2013; Tourangeau & Yan, 2007).

This threat suggests that researchers should be cautious when interpreting self-reported AI use as a direct reflection of actual behavior. Survey measures should be designed to reduce fear and ambiguity by using neutral wording, emphasizing confidentiality, separating permitted and non-permitted uses, asking about specific behaviors, and avoiding language that requires students to label their own behavior as ethical or unethical. Where possible, self-report data may also be strengthened through qualitative explanations, policy-context information, or other forms of evidence, while respecting privacy and ethical

boundaries (Podsakoff et al., 2003; Tourangeau & Yan, 2007).

4.4 Contextual and Policy Threat

The fourth threat concerns the role of institutional, disciplinary, and policy contexts. Student GenAI use does not occur in isolation. It is shaped by course rules, instructor expectations, institutional policies, disciplinary norms, technological access, language background, disability accommodations, and cultural context. Because these conditions differ across settings, the same behavior may carry different meanings in different educational environments (Chan & Hu, 2023; Kasneci et al., 2023).

For example, one instructor may encourage students to use GenAI for brainstorming but prohibit direct use in final submissions. Another instructor may prohibit all AI assistance, while a third may require students to use AI and reflect on the process. In these contexts, the same student behavior may be interpreted as appropriate learning support, academic misconduct, or required course participation. If a measure does not account for these differences, researchers may misclassify student behavior or make unfair comparisons when students differ in access to paid tools, reliable internet, institutional guidance, AI literacy, disciplinary expectations, language support, or disability-related uses of assistive technologies.

Digital divide research shows that inequalities in material access, internet skills, and patterns of use can shape what individuals are able to do online and the outcomes they obtain from technology use (van Deursen & van Dijk, 2019). If measurement instruments ignore these differences, group comparisons may reflect unequal access or unequal clarity of expectations rather than differences in motivation, ethics, or learning behavior.

A context-sensitive measurement approach should therefore include information on course expectations, awareness of institutional policies, access to technology, disciplinary setting, task type, and students' understanding of permitted use. Researchers should avoid interpreting AI use scores without considering the environment in which those scores were produced.

4.5 Interpretation and Use Threat

The fifth threat concerns how researchers, instructors, and institutions interpret and use scores from AI-related measures. Even when data are

collected carefully, validity can be threatened if the resulting scores are used to support claims that go beyond the evidence. This is especially important because student GenAI use may be interpreted in terms of learning, academic integrity, productivity, fairness, or misconduct. In the context of GenAI and assessment, Dawson et al. (2024) argue that the central issue is not only whether students cheat, but whether assessment interpretations remain valid under changing conditions of AI use.

A common interpretation problem is the overreliance on frequency. The frequency of AI use may indicate how often students report using GenAI, but it does not explain the quality, purpose, depth, or ethical status of that use. A student who uses GenAI daily for concept explanation may be engaging in productive learning support, while another may use it once to complete an assignment dishonestly. A frequency score alone cannot distinguish these cases.

Another interpretation problem is treating AI use as automatically harmful or beneficial. Some uses may support learning by helping students generate examples, receive feedback, or clarify difficult concepts, while other uses may reduce learning opportunities if students rely on AI to bypass thinking, writing, or problem solving. The educational meaning of AI use, therefore, depends on how, why, and under what conditions the tool is used (Gruenhagen et al., 2024; Kizilcec et al., 2024).

This threat also applies to institutional decision-making. If universities use weak measures to conclude that students are broadly misusing AI, they may adopt overly punitive policies. If they use weak measures to conclude that AI use is broadly beneficial, they may overlook risks related to fairness, dependency, privacy, authorship, or assessment validity. In both cases, poor measurement can lead to poor decisions.

The interpretation and use threat remind researchers that validity is not only about item design or reliability. It is also about the claims and decisions that follow from scores. For student GenAI use, responsible interpretation requires alignment between the measure, the construct, the context, and the intended use of the findings. Together, these five threats show that student GenAI use cannot be adequately measured by broad or frequency-based indicators alone. A more validity-centered approach requires researchers to clearly define the construct, examine response

processes, reduce self-report bias, account for contextual differences, and limit interpretations to claims supported by the evidence. Table 2 summarizes the five validity threats, examples from higher education, and their measurement implications.

5 From Framework to Measurement Practice

The framework is intended to help researchers move from general questions about AI adoption toward more defensible measures of student GenAI use. A stronger measure should specify the tool or category of tool, the academic task, the purpose of use, the degree of student reliance, whether the use was permitted or unclear, and the interpretation the researcher intends to make from the score. Table 1 illustrates how the framework can be used to diagnose validity problems in broad AI-use items and revise them into more precise, validity-oriented measures. The goal is not to create a single universal instrument. Rather, the goal is to show how item design can better align with construct definition, response process evidence, self-report quality, policy context, and intended score use (American Educational Research Association et al., 2014; Kane, 2013).

In addition, researchers should reduce fear-based and socially desirable responding. Since the use of GenAI may be associated with academic misconduct, students may underreport or overreport their behavior. Neutral wording, confidentiality assurances, behavior-specific questions, and attention to policy context can improve the quality of self-reported data (Krumpal, 2013; Tourangeau & Yan, 2007). Researchers should also avoid making group comparisons without considering differences in access, institutional guidance, discipline, and AI literacy.

6 Implications for Assessment, Policy, and Pedagogy

This framework has implications for assessment, policy, and pedagogy. In assessment, weak measures of AI use may lead instructors to misunderstand what student work represents. A broad report on AI use does not indicate whether students used AI for studying, editing, feedback, or completing assignments inappropriately. Better measurement can help instructors design assessments that clarify when AI support is allowed, when independent work is required, and how students should disclose AI assistance. This is important because GenAI has raised concerns

about whether assessment interpretations remain valid under changing conditions of student AI use (Apata et al., 2025; Dawson et al., 2024; Kizilcec et al., 2024).

For policy, the framework cautions against using vague survey findings to make strong claims about student misconduct or compliance. Institutions may respond too harshly if frequent AI use is automatically interpreted as cheating. They may also underestimate risk if students underreport use because they fear punishment. Clearer measurement can support policies that distinguish permitted, prohibited, uncertain, and pedagogically useful forms of AI use. This aligns with research showing that universities are developing GenAI guidelines around ethical use, authentic assessment, teaching and learning, faculty and student training, and AI literacy, while gaps remain in comprehensive policy frameworks, data privacy, and equitable access (Apata, Ajose, et al., 2026; Jin et al., 2025).

For pedagogy, valid measurement can support AI literacy instruction. If instructors know how students understand and use GenAI, they can teach clearer expectations around appropriate use, critical evaluation of AI-generated content, and responsible disclosure and the difference between using AI to support learning and using AI to replace one's own academic work. In this sense, measuring student AI use is not only a research task. It is also part of helping students engage with AI in ways that support learning rather than replace it (Chan & Hu, 2023).

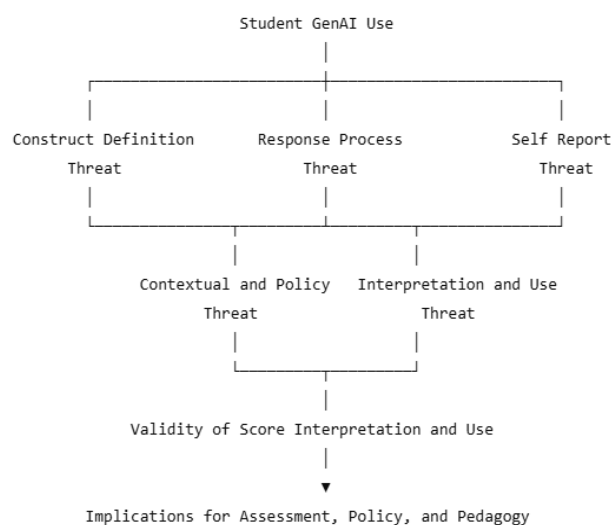
7 Conclusion

Student use of GenAI is an important issue in higher education, but it is also a difficult construct to measure. This paper argued that broad or frequency-based measures may produce weak interpretations when they fail to distinguish different types, purposes, contexts, and policy meanings of AI use. To address this concern, the paper proposed a validity-threat framework for measuring student GenAI use, organized around five threats: construct definition, response process, self-report, contextual and policy conditions, and interpretation and use. The central contribution of this paper is to position student use of GenAI as a measurement problem. Stronger measurements can support clearer construct definitions, more defensible interpretations, fairer policy, and more thoughtful pedagogical

decisions. As AI becomes increasingly embedded in assessment and pedagogy, researchers should not move too quickly from student responses to strong claims about learning, ethics, misconduct, or policy compliance. Claims about student GenAI use should be no stronger than the validity evidence supporting them.

Figure 1

A Validity Threat Framework for Measuring Student GenAI Use



Note. The framework shows that measuring student GenAI use requires attention to the definition of constructs, response processes, self-report accuracy, contextual and policy conditions, and the interpretation and use of scores. These threats shape whether evidence about student GenAI use can support claims about assessment, policy, and pedagogy.

Table 1

Illustrative Examples of Weak and Stronger Measures of Student GenAI Use

Weak item	Validity concern	Stronger alternative
How often do you use ChatGPT?	Too broad. It does not identify tool type, academic task, purpose, reliance, or policy context.	In the past four weeks, how often did you use a GenAI tool for specific academic purposes such as brainstorming, grammar editing, concept explanation, coding support, text generation, or graded work?
Have you misused AI?	Morally loaded. Students may underreport because the item	For a specific course or assignment, which forms of AI assistance were permitted,

	implies misconduct.	prohibited, or unclear?
Did AI help your learning?	Too general. It does not distinguish learning support from task completion.	For which academic tasks did GenAI help you understand course material, and for which tasks did it mainly help you produce a submission?
Do students in your program use AI ethically?	Unclear respondent knowledge and unclear ethical standards.	What AI-use expectations were communicated in the course syllabus, assignment instructions, or class discussion?

Note. These examples are illustrative and are not intended to serve as a final survey instrument. They show how validity-oriented item design can improve construct clarity, response interpretation, self-report quality, policy awareness, and score use.

Table 2

Validity Threats, Evidence Needed, and Measurement Implications for Student GenAI Use

Validity threat	Core concern	Evidence Needed	Measurement implication
Construct definition threat	The meaning of student GenAI use is unclear.	Construct map linking tool, task, purpose, reliance, and permitted use.	Define the specific type of AI use being measured.
Response process threat	Students may interpret items differently.	Cognitive interviews, pilot testing, examples, and open-ended follow-up questions.	Provide clear definitions and gather response process evidence.
Self-report threat	Students may underreport or overreport AI use.	Neutral wording, confidentiality assurance, behavior-specific items, and triangulation where ethical.	Use neutral wording and protect confidentiality.
Contextual and policy threat	Rules, expectations, and access shape responses.	Course policy, discipline, task type, access, and student awareness of expectations.	Measure policy awareness, course expectations, and access.
Interpretation and use threat	Researchers may overinterpret scores.	Explicit validity argument connecting measure, construct, context, and intended use.	Limit claims to what the evidence supports.

Note. The evidence listed in this table is illustrative rather than exhaustive. Researchers should select evidence based on the construct being measured, the context of GenAI use, and the intended interpretation and use of scores.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Apata, O. E., Ajamobe, J. O., Ajose, S. T., Oyewole, P. O., & Olaitan, G. I. (2025). The role of artificial intelligence in enhancing classroom learning: Ethical, practical, and pedagogical considerations. *Proceedings of the 2025 ASEE Gulf-Southwest Annual Conference*. American Society for Engineering Education. <https://doi.org/10.18260/1-2--55084>

Apata, O. E., Ajose, S. T., Apata, B. O., Olaitan, G. I., Oyewole, P. O., Ogunwale, O. M., Oladipo, E. T., Oyeniran, D. O., Awoyemi, I. D., Ajobiewe, J. O., Ajamobe, J. O., Appiah, I., Fakhrou, A. A., & Feyijimi, T. (2026). Artificial intelligence in higher education: A systematic review of contributions to SDG 4 (quality education) and SDG 10 (reduced inequality). *International Journal of Educational Management*, 40(1/2), 336–353. <https://doi.org/10.1108/IJEM-12-2024-0856>

Apata, O. E., Kwok, O.-M., & Ajose, S. T. (2026). The impact of ChatGPT on higher education: A systematic review of global opportunities, perceptions, and challenges. *Journal of Computer Assisted Learning*, 42, e70309. <https://doi.org/10.1002/jcal.70309>

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311. <https://doi.org/10.1093/poq/nfm006>

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, Article 43. <https://doi.org/10.1186/s41239-023-00411-8>

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, 49(7), 1005–1016. <https://doi.org/10.1080/02602938.2024.2386662>

- Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2025). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 44(11), 2518–2544. <https://doi.org/10.1080/0144929X.2024.2394886>
- Gruenhagen, J. H., Sinclair, P. M., Carroll, J.-A., Baker, P. R. A., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers and Education: Artificial Intelligence*, 7, Article 100273. <https://doi.org/10.1016/j.caeai.2024.100273>
- Jin, Y., Yan, L., Echeverria, V., Gašević, D., & Martínez-Maldonado, R. (2025). Generative AI in higher education: A global perspective of institutional adoption policies and guidelines. *Computers and Education: Artificial Intelligence*, 8, Article 100348. <https://doi.org/10.1016/j.caeai.2024.100348>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Zeivots, S., & Radulescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence*, 7, Article 100269. <https://doi.org/10.1016/j.caeai.2024.100269>
- Kofinas, A. K., Tsay, C. H.-H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, 56(6), 2522–2549. <https://doi.org/10.1111/bjet.13585>
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, 47(4), 2025–2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Li, S., Liu, J., & Dong, Q. (2025). Generative artificial intelligence-supported programming education: Effects on learning performance, self-efficacy and processes. *Australasian Journal of Educational Technology*, 41(3), 1–25. <https://doi.org/10.14742/ajet.9932>
- Lund, B. D., Mannuru, N. R., Teel, Z. A., Lee, T. H., Ortega, N. J., Simmons, S., & Ward, E. (2025). Student perceptions of AI-assisted writing and academic integrity: Ethical concerns, academic misconduct, and use of generative AI in higher education. *AI Education*, 1(1), Article 2. <https://doi.org/10.3390/aieduc1010002>
- Malik, A., Khan, M. L., Hussain, K., Qadir, J., & Tarhini, A. (2024). AI in higher education: Unveiling academicians' perspectives on teaching, research, and ethics in the age of ChatGPT. *Interactive Learning Environments*, 33(3), 2390–2406. <https://doi.org/10.1080/10494820.2024.2409407>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Qu, Y., Loo, H. E., & Wang, J. (2025). Generative artificial intelligence in higher

education: Emotional tensions and ethical declaration. *British Journal of Educational Technology*. Advance online publication. <https://doi.org/10.1111/bjet.70029>

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26(1), 100–107. <https://doi.org/10.7334/psicothema2013.256>

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883. <https://doi.org/10.1037/0033-2909.133.5.859>

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>

van Deursen, A. J. A. M., & van Dijk, J. A. G. M. (2019). The first-level digital divide shifts from inequalities in physical access to inequalities in material access. *New Media & Society*, 21(2), 354–375. <https://doi.org/10.1177/1461444818797082>

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8, 1839–1850. <https://doi.org/10.1038/s41562-024-02004-5>