

From Expert Approval to Validity Evidence: Evaluating AI-Generated Assessment Items

Chris Meador

Educational Testing Service, Princeton, NJ

[cmeador@ets.org]

Abstract

Expert approval certifies that test developers would use AI-generated items, not that the scores support valid interpretations. I extend argument-based validity with a generation inference and a five-layer evidence framework, apply it diagnostically to three NAEP–Wilbur evaluations, and propose design principles and a minimal reporting standard for the field.

1 Introduction

Large language models can now produce assessment items that are fluent, formatted, and superficially indistinguishable from professionally authored content (von Davier, 2018; 2019; Bezirhan & von Davier, 2023; Circi, Hicks, & Sikali, 2023). For programs facing chronic demand for fresh items, the appeal is clear, and adoption is moving quickly. Evaluation has lagged behind. The dominant practice for judging AI-generated items is expert review: subject-matter experts read drafts and rate their quality or usability against a rubric. This is a reasonable and necessary first step, but it is a narrow slice of what the measurement field means by validity.

The difficulty is partly perceptual. A well-formed item looks finished, and its fluency suggests that the work of validation is largely done. Yet fluency is the very property a language model is optimized to produce, and it has no bearing on whether an item measures the intended construct, discriminates among test-takers, functions equitably across groups, or supports the decisions the scores will inform. An item that an expert would happily place in a draft pool may still elicit the wrong cognition, be solvable by a construct-irrelevant shortcut, or carry difficulty no one can predict in advance. These limits hold for any item, whether human- or machine-authored. What is

distinctive about AI generation is the volume at which such items arrive and the surface polish that can discourage a second look.

This paper opens a coordinated session in which three studies evaluate Wilbur, an AI-assisted item-generation framework, across NAEP Reading, Mathematics, and Science. It argues that the move from human authoring (and from the classical, model-based automatic item generation studied elsewhere in the field) to free-form LLM generation changes where validation has to happen, and that the field needs a shared account of the evidence such systems are required to supply. (The contrast here is with the automatic-item-generation literature; NAEP item development itself has been human-authored, not model-based.) I make five contributions. First, I explain why LLM generation differs from earlier automatic item generation in a way that matters for validity: the disappearance of the item model removes the inheritance of warranted properties. Second, I extend the argument-based validity framework with an explicit generation inference. Third, I propose a five-layer framework of validity evidence for AI-generated items, with each layer tied to its methods, its AI-specific threats, and its relative cost, and mapped to the established sources of validity evidence (Table 1). Fourth, I apply the framework to the session's NAEP evaluations, using it diagnostically to locate what those studies establish and what they leave open. Fifth, I draw out implications for designing generation systems that emit evidence, and propose a reporting agenda, including a minimal reporting standard, for studies of AI-generated items.

2 Background

2.1 Validity as an Argument

Modern validity theory treats validity not as a property an item or test possesses but as the degree

to which evidence and theory support the interpretations and uses of scores (Messick, 1989; AERA, APA, & NCME, 2014). An item is never valid in the abstract; it is valid, or not, relative to a specific claim about what its scores mean and how they will be used. The Standards organize the supporting evidence into sources based on test content, response processes, internal structure, relations to other variables, and the consequences of testing, and they treat fairness as a fundamental and pervasive concern throughout.

Kane's argument-based approach makes the logic explicit (Kane, 2006, 2013). Validation proceeds by laying out an interpretation/use argument as a chain of inferences, typically scoring, generalization, extrapolation, and implication, and then assembling evidence to back each inference and to rebut plausible threats to it. The strength of the overall argument is limited by its weakest link. This framing shifts attention from "is this item good?" to "which inferences are we asking these items to support, and what backs each one?"

2.2 From Item Models to Free Generation

Automatic item generation is not new. For two decades, programs have generated items from explicit cognitive and item models: a template, or item model, specifies the radicals that drive an item's psychometric properties and the incidentals that may vary without consequence (Irvine & Kyllonen, 2002; Bejar, 2002; Embretson, 1998; Gierl & Haladyna, 2013). The validation logic of model-based generation is straightforward: one validates the model, and individual items, as instances of the model, inherit much of its warrant. Because the features that drive difficulty are controlled by design, item difficulty is partly predictable a priori, and large families of items can be produced with known and comparable properties (Gierl, Lai, & Turner, 2012).

LLM generation breaks this logic. The transition began with neural approaches that learned to produce item text directly instead of instantiating a parameterized model (von Davier, 2018, 2019), and it has accelerated with general-purpose LLMs that generate items from natural-language prompts (Bezirhan & von Davier, 2023; Circi, Hicks, & Sikali, 2023). There is typically no explicit item model, no controlled separation of radicals from incidentals, and no parameterized family from which an item descends. The item is produced free-

form by an opaque process conditioned on a prompt. Nothing about the output's psychometric behavior is guaranteed by construction; difficulty, discrimination, and dimensionality are unknown until observed. The consequence for validation is fundamental: items can no longer inherit warranted properties from a validated model. The unit that must be validated moves either to the individual output or to the generation process treated as a measurement instrument whose products require ongoing evidence. Expert review of drafts, the field's current default, fully addresses neither: it can establish content evidence for an individual draft, but it does not validate the generation process as an instrument or supply the empirical evidence the later inferences require.

2.3 The System Under Study: Wilbur

Wilbur is an ETS-developed, AI-enabled framework for assessment item development; this paper focuses on its application to the NAEP item development lifecycle within a human-in-the-loop workflow. Wilbur does not produce finished items, but instead it drafts candidates that subject-matter experts review and revise. A user specifies generation parameters (grade, content target, item type, and framework requirements) in a structured prompt aligned to NAEP specifications; the system generates a batch of draft items; and reviewers iterate with the model for a set number of structured feedback loops before items are scored against an a priori rubric. The system records the model version, prompt, and generation path for each item. I introduce Wilbur here as the system the companion Reading, Mathematics, and Science studies evaluate. Its design choices reappear in Section 7 as a concrete instance of the design and process evidence (Layer 1) that the framework calls for, so that the generator can be treated as an instrument whose products require ongoing evidence.

3 What Expert Review Establishes, and What It Doesn't

Expert review of AI-generated items is a form of evidence based on test content: experts judge whether items are aligned to the targeted construct and framework, are accurate, are grade-appropriate, and are free of obvious bias (AERA, APA, & NCME, 2014; Haladyna & Rodriguez, 2013). This corresponds to one of the five sources of validity evidence, and it is important. Recent

studies in this venue confirm both its value and its ceiling: experts judging AI-generated clinical items found many usable as helpful starting points but few turnkey (Li et al., 2025; Afrin et al., 2025). Two limitations are easy to overlook.

First, content evidence is only one source. Establishing that experts would accept an item says little about whether students engage the intended cognition (response processes), whether the item behaves as expected alongside others (internal structure), whether it functions equitably across groups (fairness), or whether scores built from such items support the intended decisions (relations and consequences). Because a finished-looking item reads as complete, reviewers may treat expert acceptance as proof of sufficiency, when it establishes only one part of the case.

Second, content evidence is only as trustworthy as the process that produces it. Expert judgments support a claim about content only if raters agree at a level that can be quantified and if the rubric they apply is itself a defensible instrument. When reviewer agreement is reported only as raw percent agreement instead of a chance-corrected coefficient, and the rubric's own reliability and the basis for its cut scores are not established, the content claim is under-warranted: not necessarily false, but unsupported. This is a common gap, and one of the cheapest to close. Agreement is necessary but not sufficient; recent work cautions against treating rater agreement as ground truth in its own right and argues for pairing it with evidence that the judgments track what matters (Thomas, Borchers, & Koedinger, 2025).

There is also a threat specific to AI-generated content. Language models are optimized to produce text that reads as plausible and well-formed, the very surface on which content review operates (Bender et al., 2021). Fluent, confidently written items can induce a kind of automation bias in reviewers, who may extend to a machine-authored draft a benefit of the doubt earned by human expertise (Skitka, Mosier, & Burdick, 1999). The property that makes AI items attractive to generate is also what makes them difficult to screen by inspection alone.

These limits are visible in current practice. The NAEP item-generation evaluations that anchor this session found that experts accepted a substantial share of generated items as usable drafts while judging none operationally ready, and they documented a consistent gradient: judged quality

declined as the cognitive demand of the targeted construct rose, from lower-order to higher-order reading comprehension, from procedural mathematics to items targeting mathematical practices such as justifying and proving, and from single-dimension to multidimensional science items aligned to both a disciplinary core idea and a science-and-engineering practice. That gradient is informative. It marks the ceiling of what generation plus expert review currently achieves, and it points to where deeper evidence is most needed. Higher-order constructs are where construct-irrelevant solvability and distractor infidelity are most likely, and where content inspection (as for any item, but with higher stakes here) is least able to certify that an item elicits the intended reasoning.

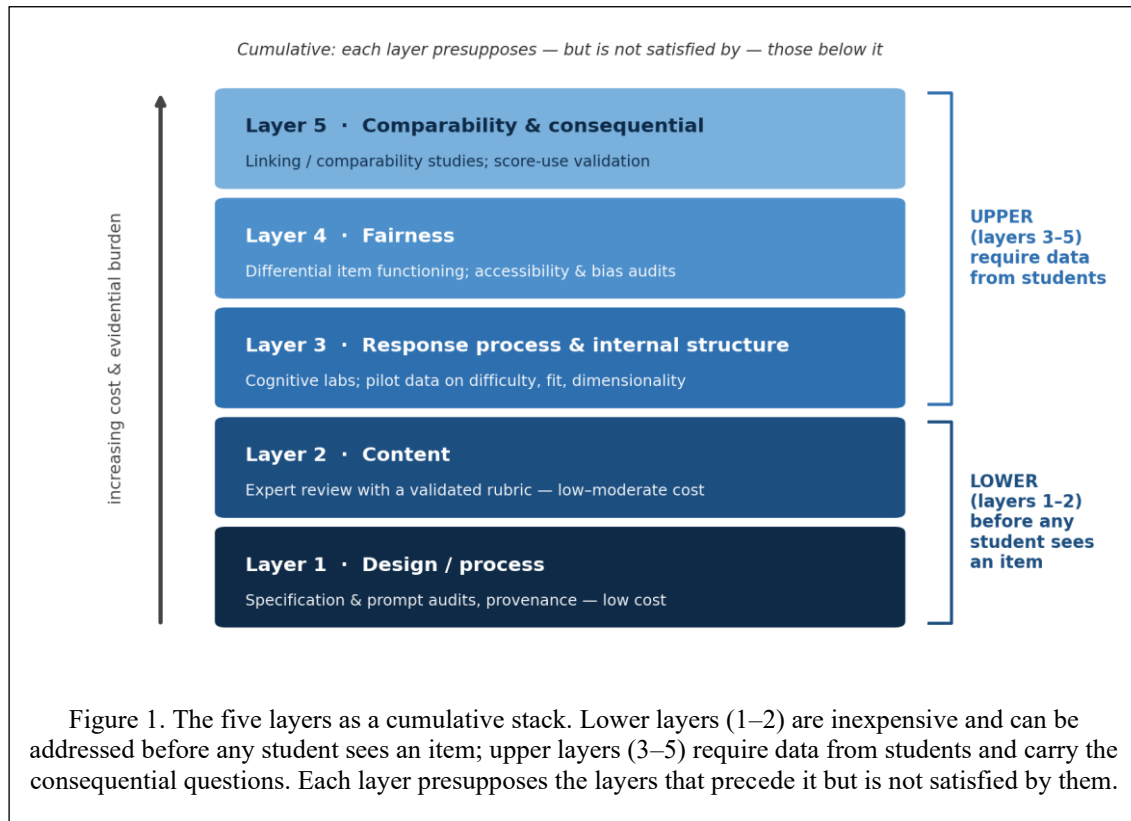
4 A Layered Framework of Evidence for AI-Generated Items

4.1 The Generation Inference

I propose adding an explicit inference to the interpretation/use argument for any assessment built on generated items: the generation inference, which links a target specification (the construct, framework objective, and intended evidence) to the item the system produces. Its warrant is the claim that the generation process reliably yields items that elicit the intended evidence about the targeted construct. Its threats include misalignment between specification and output, the insertion of construct-irrelevant cues that make items solvable without the target competence, distractors that do not correspond to genuine student misconceptions, contamination from training data, and homogenization across a generated set. The generation inference is logically prior to scoring (Kane, 2013): if it fails, no amount of careful scoring or scaling downstream can recover a valid interpretation. Making it explicit forces the question of what supports it, so that fluency cannot quietly stand in for evidence.

4.2 Five Layers of Evidence

I organize the evidence an AI generator must supply into five layers, ordered roughly from least to most costly and from process to consequence. The layers are cumulative, not alternative: higher layers presuppose the lower ones but are not satisfied by them (Figure 1).



Layer	Claim warranted	Methods and data	Characteristic AI-specific threats	Cost	Standards source
1. Design / process	Generation embodies the intended assessment argument	Specification and prompt audits; documented evidence model per item; framework traceability; provenance logs	Specification underdetermines the construct; prompt/planner drift; undocumented process	Low	Content (at design)
2. Content	Items are aligned, accurate, fair on their face, grade-appropriate	Expert review with a validated rubric and reported inter-rater reliability	Fluency/automation bias; rater unreliability; invalid or arbitrary rubric	Low–mod.	Content
3. Response process & internal structure	Items elicit intended cognition and behave as expected	Cognitive labs/think-alouds; pilot data on difficulty, discrimination, distractor functioning, dimensionality, fit	Construct-irrelevant solvability; unpredictable difficulty; non-functioning distractors	High	Response processes; internal structure
4. Fairness	Items function equitably across groups	Differential item functioning; accessibility review; bias audits; subgroup performance	Social bias in training data; differential contamination; accessibility defects	High	Fairness; relations
5. Comparability & consequential	AI items support the same uses as human-authored items, with acceptable consequences	Comparability/linking studies; score-use validation; monitoring of consequences	Construct shift at scale; homogenization; over-reliance eroding human review	High	Relations; consequences

Table 1. Five layers of validity evidence for AI-generated items, with methods, characteristic AI-specific threats, relative cost, and mapping to the sources of evidence in the Standards (AERA, APA, & NCME, 2014).

The first two layers are inexpensive and can be addressed before any student sees an item, and they are where current practice lives. The upper three require empirical data from real test-takers and are where the consequential validity questions reside. The mapping to the Standards' sources is deliberate: the framework reorganizes established validity theory for the AI-generation setting; it is not a new or competing standard. Its contribution is to make each layer's methods, its characteristic threats, and its cost explicit, so that a program can see which inferences it has backed and which it has merely assumed. Several contemporaneous efforts pursue related goals, such as frameworks for dynamic validity under human–AI collaboration (Feyijimi et al., 2025), and I position the present framework as the evidentiary backbone for the studies in this session.

5 AI-Specific Threats and the Scale Paradox

Several threats recur across the layers and deserve to be named as a set, because they are characteristic of LLM generation, not of item writing in general.

Construct-irrelevant solvability. Generated items may be answerable through surface heuristics, response-option asymmetries, or verbal cues, allowing test-takers to succeed without the target competence. Trained item writers learn to suppress these flaws through item-writing guidelines; a generator honors them only insofar as the prompt encodes such rules, so in their absence it tends to reproduce the surface regularities of its training data. This inflates apparent performance and corrupts the response-process inference.

Misconception infidelity. Plausible-looking distractors are easy to generate; distractors grounded in the specific misconceptions students hold are not. Distractors that look reasonable to an adult but do not reflect real error patterns degrade discrimination and license incorrect diagnostic inferences.

Contamination and leakage. Models trained on public corpora may reproduce or closely paraphrase released items or generate content that advantaged students have effectively seen. The result is both a validity and a fairness threat that inspection cannot detect, and it is documented at length in the evaluation of LLMs themselves (Xu et al., 2024; Cheng et al., 2025).

Homogenization. Generators tend toward the modal and the repetitive (Bender et al., 2021). Sets

of items that cluster around a few templates or contexts narrow construct coverage and can inflate internal-consistency reliability in ways that do not generalize. This is an internal-structure problem, not the reliability gain it appears to be.

Difficulty opacity. Unlike model-based generation, free generation offers no a priori handle on difficulty, so banks cannot be assembled to target information functions without empirical calibration (von Davier, 2018; Circi, Hicks, & Sikali, 2023). Emerging work on predicting item parameters from item text aims to narrow this gap (Han et al., 2025).

Content and computation errors. Even fluent items can contain factual or computational errors, a documented failure mode for LLMs on quantitative tasks (Zhang & Graf, 2025).

Fluency and automation bias. The readability of generated items can suppress the very scrutiny they require (Skitka, Mosier, & Burdick, 1999).

These threats interact with a structural tension I term the scale paradox. The value proposition of AI generation is volume: many items, quickly. But the evidence that matters most, in layers 3 through 5, is gathered per item, from students, at substantial cost, and it does not scale at the same rate as generation. Left unmanaged, this produces validity debt: item banks accumulating far faster than the warrants that would justify their use. The discipline's analogue to paying down such debt is to validate the generator as an instrument, field-testing representative samples and monitoring it continuously instead of relying on a one-time check, while still requiring that any individual item carry its own empirical evidence before operational use, and being explicit about which items in a bank are backed by which layers of evidence.

6 Locating Current Practice: The NAEP Evaluations

The framework is meant to be used diagnostically, and the three evaluations in this session offer a concrete demonstration. Each applied a structured expert-review rubric to a substantial sample of Wilbur-generated items, and together they supply layer-2 content evidence across three subjects while highlighting what is distinctive about each. The Reading study (DiMeglio) is the only passage-based subject; it evaluates 450 items across grades 4, 8, and 12, contrasts constrained and free-rein prompting, and traces performance across comprehension targets and item categories, with

acceptance highest for Locate and Recall items and lowest for Language Structures and Vocabulary items. The Mathematics study (Van Ness, Parker, & Minchuk) foregrounds the contrast between content-aligned items and items targeting NAEP Mathematical Practices and is the only study to transition models mid-evaluation (GPT-4 to GPT-5.2). The Science study (Safar, Price, & Ackley) raises the bar to multidimensional items aligned to both a disciplinary core idea and a science-and-engineering practice and is the only study to benchmark Wilbur against external sources: outside item writers and the NGSS Item Authoring Tool. Across all three, judged quality declined as construct complexity rose, and no items were judged operationally ready. All three report reviewer agreement on accept/reject decisions (percent agreement of 58% to 80%), though none reports a chance-corrected reliability coefficient.

The framework also makes the open questions precise, without disparaging the work. The layer-2 evidence is, in places, under-warranted: reviewer agreement is reported only as raw percent agreement instead of a chance-corrected coefficient, and the bespoke rubrics' properties and cut scores are not independently justified. Both are inexpensive to remedy, often from data already collected. Above layer 2, the studies are silent by design: there is no response-process or internal-structure evidence (no student response data, and hence no empirical difficulty, discrimination, or dimensionality), no fairness evidence (no differential item functioning analysis), and no comparability or consequential evidence relating AI-generated items to human-authored ones. The observed complexity gradient is best read as a signal that, for higher-order constructs in particular, the missing upper-layer evidence is not optional: content review alone cannot certify that a hard item elicits hard thinking. The evaluations did the first layer well. The framework names the layers still to come, and in doing so turns a set of parallel quality studies into the opening move of a validity argument. The three talks that follow develop it subject by subject.

7 Designing Systems to Produce Evidence

If validity evidence is expensive and does not scale, one response is to design generation systems so that they emit evidence as a byproduct instead of leaving it to be reconstructed afterward. Several

design choices, illustrated by Wilbur and by the multi-stage generation architectures now emerging, can lower the cost of the lower layers and instrument the system for the upper ones.

A generation pipeline that begins with an explicit planning stage, one that analyzes the objective and intended evidence, enumerates target misconceptions, and records these before any drafting, is in effect executing an evidence model in the sense of Evidence-Centered Design (Mislevy, Steinberg, & Almond, 2003). The artifact it produces is layer-1 design evidence, captured automatically and traceably. Deterministic verification stages that check computational correctness, readability bands, and structural constraints log objective quality signals that constitute an automatable slice of content evidence, reserving scarce expert attention for judgments only humans can make; this is especially valuable given documented LLM weaknesses on computation (Zhang & Graf, 2025). Provenance metadata that records the model version, prompt, and generation path for each item, as Wilbur does, supports contamination audits and reproducibility, echoing transparency-documentation practices from machine learning more broadly (Mitchell et al., 2019; Gebu et al., 2021). Automated framework-alignment checks (Bezirhan & von Davier, 2025) and built-in diversity controls that counter homogenization address content and internal-structure concerns during generation instead of after a bank has been assembled. Some early evidence at the upper layers can even be approximated before piloting: LLM-simulated responses and learned item-parameter predictors offer provisional signals about difficulty and functioning (Han et al., 2025; Demirkaya, Wei, & Johnson, 2025).

Two cautions are essential. First, evidence a system emits about its own process is still only layer-1 and layer-2 evidence and simulated or predicted signals are provisional; they reduce costs at the lower layers but cannot substitute for empirical response data from students. A system cannot certify its own items' difficulty or fairness. Second, instrumenting a generator for evidence must not be mistaken for the evidence itself; the value of these design choices is that they make claims checkable, and the checking still must happen.

8 A Reporting and Research Agenda

The framework implies a sequence of commitments, ordered by cost and urgency. In the near term, and at low cost, studies of AI-generated items should report the reliability of expert judgments and validate the rubrics those judgments rest on, document generation provenance, and adopt a minimal reporting standard so that readers can locate a study on the evidence framework. Reporting guidelines have improved transparency in other fields, where model cards and datasheets are useful precedents (Mitchell et al., 2019; Gebru et al., 2021), and AI-generated assessment is overdue for an equivalent. I propose that such a study report, at minimum:

- the model and version, and the generation process (single-shot or multi-stage, with prompts or specifications);
- the sample of items and the targets they address;
- rater qualifications, training, and quantified inter-rater reliability;
- the rubric and the evidence for its reliability and cut scores;
- and, explicitly, which layers of validity evidence are claimed, and which are not.

The final item is the most important: a study that states plainly that it provides content evidence only, and names what remains to be shown, is more useful, and more credible, than one that implies more than it shows. In the medium term, programs should move up the framework: embed generated items in pilots to obtain response-process and internal-structure evidence; conduct differential item functioning and accessibility analyses for fairness (Holland & Wainer, 1993; Gallegos et al., 2024); and run comparability studies that test directly whether AI-generated and human-authored items support equivalent interpretations and uses (cf. Sonnleitner, 2025). At the field level, shared benchmarks and evidentiary expectations would let results accumulate across programs, and treating the generator as an instrument under ongoing validation, not a tool cleared once, would keep validity evidence current as output grows.

9 Conclusion

Expert approval starts a validity argument; it does not finish one. The shift from model-based generation to free-form LLM generation removes the inheritance that once let individual items borrow a model's warrant and so returns the burden of evidence to the output and the process that produced it. The framework offered here names what that burden consists of: a generation inference that must be backed, and five layers of evidence (design; content; response process and internal structure; fairness; and comparability and consequences), each with its own methods, threats, and costs. Applied to current practice, including the NAEP evaluations in this session, the framework shows a field operating responsibly at its lower layers while the consequential questions wait at the upper ones. As generation scales, the central task for educational measurement is to build the higher-layer evidence as quickly as the generator builds items: by designing systems that produce evidence, and by reporting honestly about which inferences are backed and which are still missing.

10 Limitations

This is a conceptual contribution: the framework synthesizes and reorganizes established validity theory for a new setting, and it has not itself been validated in use. The layers are a heuristic ordering, not a strict sequence, and real programs may gather evidence out of order or in parallel; the cost designations are relative and context-dependent. The framework describes the evidence that would warrant operational use and does not prescribe thresholds, which remain a matter of program-specific argument.

References

- Afrin, T., Ha, L. A., Yaneva, V., Evanini, K., Go, S., DeRuchie, K., & Heilig, M. (2025). Toward automated evaluation of AI-generated item drafts in clinical assessment. In *Proceedings of AIME-Con: Full papers*. NCME.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen

- (Eds.), Item generation for test development (pp. 199–217). Lawrence Erlbaum.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of FAccT '21 (pp. 610–623). ACM.
- Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5, Article 100161.
- Bezirhan, U., & von Davier, M. (2025). AI-based classification of TIMSS items for framework alignment. In Proceedings of AIME-Con: Full papers. NCME.
- Cheng, Y., Chang, Y., & Wu, Y. (2025). A survey on data contamination for large language models. arXiv. <https://arxiv.org/abs/2502.14425>
- Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8, Article 858273.
- Demirkaya, O., Wei, H.-R., & Johnson, E. (2025). Simulating rating scale responses with LLMs for early-stage item evaluation. In Proceedings of AIME-Con: Full papers. NCME.
- DiMeglio, C. (2026). Evaluating AI-generated NAEP items: The assessment developer's perspective—Reading. Paper submitted for AIME-Con.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380–396.
- Feyijimi, T., Oyeniran, D. O., Apata, O., Makinde, H. S., Adegoke, H. O., Ajamobe, J., & Dadzie, J. (2025). Augmented measurement framework for dynamic validity and reciprocal human-AI collaboration in assessment. In Proceedings of AIME-Con: Full papers. NCME.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097–1179.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
- Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.
- Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8), 757–765.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
- Han, S., Rijmen, F., Boykin, A. A., & Lottridge, S. (2025). Leveraging fine-tuned large language models in item parameter prediction. In Proceedings of AIME-Con: Full papers. NCME.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum.
- Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Lawrence Erlbaum.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Li, J., Rezayi, S., Baldwin, P., Harik, P., & Yaneva, V. (2025). Towards reliable generation of clinical chart items: A counterfactual reasoning approach with large language models. In Proceedings of AIME-Con: Full papers. NCME.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of FAccT '19 (pp. 220–229). ACM.
- National Assessment Governing Board. (2021). *Reading framework for the 2026 NAEP*.
- National Assessment Governing Board. (2021). *Mathematics framework for the 2026 NAEP*.
- Safar, C., Price, M., & Ackley, J. (2026). *Evaluating AI-generated NAEP items: The assessment developer's perspective—Science*. Paper submitted for AIME-Con.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
- Sonnleitner, P. (2025). Evaluating AI-generated vs. human-written reading comprehension passages:

An expert SWOT analysis and comparative study for an educational large-scale assessment. *Large-scale Assessments in Education*, 13, Article 20. <https://doi.org/10.1186/s40536-025-00255-w>

Thomas, D. R., Borchers, C., & Koedinger, K. (2025). Beyond agreement: Rethinking ground truth in educational AI annotation. In *Proceedings of AIME-Con: Full papers*. NCME.

Van Ness, C., Parker, K., & Minchuk, I. (2026). *Evaluating AI-generated NAEP items: The assessment developer's perspective—Mathematics*. Paper submitted for AIME-Con.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, 83(4), 847–857.

von Davier, M. (2019). Training Optimus Prime, M.D.: Generating medical certification items by fine-tuning OpenAI's GPT-2 transformer model. *arXiv*. <https://arxiv.org/abs/1908.08594>

Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark data contamination of large language models: A survey. *arXiv*. <https://arxiv.org/abs/2406.04244>

Zhang, L., & Graf, E. (2025). Mathematical computation and reasoning errors by large language models. In *Proceedings of AIME-Con: Full papers*. NCME.