

Evaluating an AI-Item Generation Tool for NAEP Mathematics

Cheryl Van Ness, Karen Parker, Ilona Minchuk

ETS Research Institute, Princeton, NJ

[cvanness@ets.org]

Abstract

This study evaluates the effectiveness of an AI-based system (Wilbur) for generating NAEP mathematics items. Results from this investigation indicate moderate success for content-focused items but substantial challenges for items intended to assess the NAEP Mathematical Practices. Transitioning from GPT-4 to GPT-5.2 improved item quality, though human revision remained essential.

1 Background

Generative artificial intelligence (AI), and particularly large language models (LLMs), have become increasingly common tools used in educational assessment, especially for supporting the development of draft assessment items. Recent research shows that LLMs can generate mathematics items that align with targeted content objectives, suggesting potential for use in large-scale assessment development (Meissner et al., 2024). At the same time, mathematics-specific studies identify consistent limitations. Generated items may include incomplete or incorrect solution processes, and LLMs often struggle to produce problems that require the integration of multiple mathematical skills or extended reasoning (Meissner et al., 2024; Shah et al., 2025). More broadly, reviews of AI-based item generation note that while these systems can generate linguistically well-formed items efficiently, many approaches lack strong grounding in measurement principles such as construct validity and alignment to defined frameworks (Tan et al., 2025).

To investigate the extent to which an AI-based system could support the generation of high-quality items for a large-scale mathematics assessment, this study employed Wilbur—an AI-enabled item generation tool that leverages multiple LLMs to scale and enhance various stages

of the assessment development lifecycle—to generate items aligned with the measurement constructs of the NAEP Mathematics assessments. These items were evaluated on construct alignment, accuracy and defensibility of keys, plausibility of distractors, and adherence to fairness and editorial guidelines.

The NAEP mathematics context provides an important test case for AI-based item generation. The 2026 Mathematics Framework for the National Assessment of Educational Progress describes NAEP Mathematics as an assessment of students' understanding of mathematical content and their engagement in mathematical practices, organized around five content areas—Number Properties and Operations, Measurement, Geometry, Data Analysis, Statistics, and Probability, and Algebra—and five NAEP Mathematical Practices (National Assessment Governing Board, 2021). These practices describe key dimensions of doing mathematics. Representing involves recognizing, using, creating, interpreting, or translating among representations. Abstracting and Generalizing involves decontextualizing, identifying commonalities, and extending one's reasoning to a broader domain. Justifying and Proving involves creating, evaluating, showing, or refuting generalized mathematical claims. Mathematical Modeling involves making sense of a scenario, identifying a problem to be solved, mathematizing it, applying the mathematization to reach a solution, and checking the viability of the solution. Collaborative Mathematics involves the social enterprise of doing mathematics with others through discussion and collaborative problem solving whereby ideas are offered, debated, connected, and built upon toward a solution and shared understanding. Because these practices are assessed alongside content, NAEP mathematics items must capture both what students know and how they use that knowledge. This requires integrating reasoning,

interpretation, and multiple aspects of construct alignment within a single task, making practice-aligned items potentially more difficult targets for AI-based generation than content objectives alone.

The human process of developing NAEP mathematics items reflects this multidimensional framework. Subject matter experts identify targeted constructs that include both content objectives and Mathematical Practices, and human item writers draft items aligned to detailed specifications. These items then move through an iterative development cycle, including content, fairness, editorial, committee, and client reviews, with revisions at each stage to ensure accuracy, clarity, and alignment. Each item is intentionally developed to measure a specific combination of content, practice, and item type, and is evaluated as a unique representation of that construct within the assessment.

This approach differs substantially from much of the prior research on automated item generation, which has largely relied on template- or model-based methods (Sonnleitner et al., 2025; Tan et al., 2025). In these approaches, item features are defined in advance using structured item models, allowing systematic variation across items and enabling control over characteristics such as difficulty and cognitive demand. For example, cognitive item modeling approaches demonstrate that item properties can be driven by predefined factors, supporting scalable generation within a controlled design space (Sonnleitner et al., 2025). While these methods support strong construct control, they rely on predefined structures and are typically optimized for well-defined content domains. In contrast, NAEP mathematics item development does not rely on template-based generation. In a given round of development, each item is uniquely constructed to reflect a distinct intersection of content, practice, and format, rather than being derived from a common model or variation of a shared structure. For this reason, the evaluation process in the present study emphasized the production of workable first drafts, where the ideation and creative development process can be especially time intensive. Thus, this study addresses a gap in the research by examining the potential of AI-based item generation in a non-template-driven context.

Rather than generating multiple items from a shared structure, the task requires producing pairs of items that satisfy unique measurement targets

aligned to the NAEP framework. This represents a more complex and less constrained design space than those typically examined in prior research. By evaluating AI-generated items within this framework, the present study investigates whether LLMs can support item development in a context that requires flexible generation across heterogeneous constructs, including those that depend on reasoning, explanation, and the integration of multiple mathematical processes.

2 Key Questions

The present paper focuses on the evaluation of Wilbur's capabilities in generating assessment items for NAEP Mathematics and addresses four primary questions:

1. Can Wilbur generate mathematics items that reliably meet NAEP expectations for quality, validity, and fairness?
2. Does item quality vary by grade level, content area, and NAEP Mathematical Practice?
3. Which features of the NAEP Mathematics assessment create challenges for AI-enabled item generation?
4. How does large language model choice affect the quality of generated items?

Within the scope of this study, item quality referred specifically to expert-assessed usability of draft items, including the extent to which an item aligned with the intended 2026 assessment framework targets (i.e., objectives and mathematical practices) and met program-defined expectations for mathematical correctness, grade-level appropriateness, clarity, fairness, accessibility, and style. Because no student-response data were collected, the study did not address operational characteristics such as item difficulty, discrimination, or validity as reflected in empirical performance. Rather, the purpose was to evaluate whether Wilbur could generate draft items of sufficient quality to warrant continued human refinement within the existing NAEP Mathematics review process.

3 Samples

The sample for this study was chosen to support the evaluation of items across a range of objectives, content areas, and Mathematical Practices. It consisted of 60 Wilbur-generated items intended to assess grades 4 and 8 NAEP mathematics measurement targets. Although NAEP mathematics also includes grade 12, that content is substantially more complex, so grade 12 items were not included in this study. Focusing on grades 4 and 8 allowed the evaluation to examine the NAEP Mathematical Practices without the added challenge of advanced high school content. This design was intended to reduce the number of feedback iterations that might otherwise have been needed to address both content complexity and practice demands at the same time. The item type was limited to single-select multiple choice (SSMC) because this is an item type that is commonly used in NAEP Mathematics assessments and one that Wilbur has previously supported.

The 60 items were evenly divided across grades 4 and 8 and distributed across the five NAEP mathematics content areas. For each objective, six items were generated: two non-practice-aligned (“Other”) items and four practice-aligned items, of which two targeted one NAEP Mathematical Practice and two targeted another. To reflect the typical mathematics item-writing process, in which the practice is specified but the subskill within the practice is selected by the item writer, Wilbur was prompted to choose the skill within the practice for each item.

In total, the generated items covered approximately 7% of objectives at grades 4 and 8. Objectives used to generate the items were selected to reflect clear, straightforward wording with minimal need for clarification, and objective-practice pairings were chosen to emphasize natural applications rather than forced combinations. Because each item aligned with a unique measurement target, items were generated in separate sessions rather than through batch generation.

4 Methods

Item generation followed an iterative process involving structured prompts aligned to NAEP specifications and up to three rounds of feedback and revisions. In some cases, earlier iterations were

retained for evaluation if they were better aligned to intended constructs.

A key feature of the methodology was the transition from GPT-4 to GPT-5.2 during the evaluation. This shift was implemented in response to low acceptance rates under GPT-4, particularly for practice-aligned items. Of the 60 items, 24 were generated using GPT-4 and 36 using GPT-5.2.

Items were evaluated using a standardized rubric designed for SSMC mathematics items. The rubric was adapted from established NAEP item review criteria and served as the instrument for the expert reviews in the study. Evaluation included 12 criteria separated into the following four categories:

- Alignment with targeted constructs (i.e., objectives and mathematical practices)
- Content accuracy
- Key accuracy and the quality of distractors and rationales
- Adherence to NAEP style and accessibility guidelines, including the absence of bias or sensitivity concerns

Each item was scored on 12 criteria using a three-point scale, with total scores mapped to four recommendation categories: Acceptable with No Revisions, Acceptable with Minor Revisions, Acceptable with Major Revisions, or Rejected. One criterion evaluated the usability of image descriptions included with the item. Since that criterion did not apply to all items, cut scores varied depending on whether an item included an image description. The minimum acceptable score reflected an average of 2 points per applicable criterion; items below that threshold were rejected. Although no single criterion functioned as a strict gatekeeper, construct alignment and content accuracy carried more weight in the rubric than stylistic guidelines.

Items were reviewed independently by at least two mathematics subject-matter experts. To ensure rater accuracy and consistency, some items were triple- and quadruple-scored. After independent review, all items were discussed to reach a final consensus rating. The findings presented in this paper are drawn from those consensus ratings and should be understood as results of expert content review.

5 Results¹

5.1 Overall Results

Overall, 37% of items were accepted with minor revisions, 47% were accepted with major revisions, and 17% were rejected. No items were accepted with no revisions.

5.2 Results by Grade Level

Item quality differed by grade. Grade 8 items were more likely to be accepted with minor revisions and less likely to be rejected than grade 4 items, although outcomes were similar across grades.

Grade	Accept Minor	Accept Major	Reject
Grade 4	30%	47%	23%
Grade 8	43%	47%	10%

Table 1: Acceptance rates by grade

5.3 Results by Content Area

Variation across content areas was minimal. Algebra items showed somewhat stronger performance, while Geometry items required more frequent major revisions despite having no rejections. Other areas displayed mixed outcomes, suggesting that content domain alone did not strongly determine item quality. Although the visual aspects of Geometry might have some effect on the quality of the items generated by Wilbur, these results indicate that, in general, factors other than content area more significantly affected item acceptance rates.

Grade	Accept Minor	Accept Major	Reject
Number Properties and Operations	42%	33%	25%
Measurement	33%	42%	25%
Geometry	25%	75%	0%
Data Analysis, Statistics, and Probability	42%	33%	25%
Algebra	42%	50%	8%

Table 2: Acceptance rates by content area

5.4 Results by Mathematical Practice

The most pronounced differences in item quality were observed between practice-aligned and non-practice items. Non-practice items performed substantially better, with a majority accepted under the category of minor revisions and none rejected. By contrast, practice-aligned items accounted for all rejections and were more likely to require major revisions.

Practice	Accept Minor	Accept Major	Reject
Representing	25%	63%	13%
Abstracting and Generalizing	25%	38%	38%
Justifying and Proving	0%	88%	13%
Mathematical Modeling	50%	38%	13%
Collaborative Mathematics	13%	38%	50%
Other	65%	35%	0%

Table 3: Acceptance rates by Mathematical Practice

Performance also varied across specific practices. Items targeting Justifying and Proving and Collaborative Mathematics showed particularly weak outcomes, while Mathematical Modeling exhibited somewhat stronger performance relative to other practices.

These results indicate that constructs requiring reasoning and explanation are substantially more difficult for AI systems to operationalize.

5.5 GPT-4 vs. GPT-5.2 Results

The transition from GPT-4 to GPT-5.2 resulted in notable improvements in item quality, particularly in reducing rejection rates and increasing acceptance rates.

Model	Accept Minor	Accept Major	Reject
GPT-4	13%	38%	50%
GPT-5.2	29%	63%	8%

Table 4: Acceptance rates by GPT type

Rejection rates decreased substantially under GPT-5.2, while acceptance with minor revisions increased. These improvements were especially

¹ Percentages may not sum to 100% due to rounding.

notable for practice-aligned items, which had previously demonstrated high failure rates.

5.6 Reviewer Agreement

Reviewer agreement was examined using reviewers' independent pre-consensus ratings. Agreement was evaluated based on "accept" and "reject" categories during the double-scoring process.

Across all 60 items, reviewers matched 48 out of 60 items (80%). Agreement was slightly higher for grade 4 (25 out of 30 items, 83%) than for grade 8 (23 out of 30 items, 77%). For non-practice ("Other") items, reviewers agreed on accept versus reject classifications for all 20 out of 20 items (100%). For practice-aligned items, reviewers agreed on 28 out of 40 items (70%). These items accounted for all four of the accept-reject disagreements.

Overall, these results provide descriptive evidence of reviewer consistency across the 60 generated items.

6 Discussion

This study examined the extent to which large language models can support the generation of mathematics assessment items within the NAEP framework, a context defined by multidimensional constructs and non-template-based item development. Across the evaluated items, results indicate that AI-generated content demonstrates clear potential for supporting early-stage item development, but also reveals consistent limitations, particularly for items targeting the NAEP Mathematical Practices.

Overall, while most items showed promise, the majority required substantive refinement before meeting NAEP standards for usable draft items. Differences by construct dimension were especially pronounced, with the quality of the non-practice items being higher than that of practice-aligned items, demonstrating that item generation success varies systematically based on the complexity of the construct being assessed, rather than uniformly across content or grade level.

This pattern is consistent with prior research on AI-generated mathematics assessment items. Studies have shown that LLMs are effective at generating items aligned to content objectives and cognitive levels, but often produce incomplete reasoning, inaccuracies, or insufficient complexity in solution processes (Meissner et al., 2024). Other

work highlights challenges in generating problems that require multi-step reasoning or the integration of multiple mathematical skills, which are essential components of higher-order mathematical thinking (Shah et al., 2025). The present findings extend this literature by situating these challenges within a formal and public assessment framework, showing that these limitations are most evident in items aligned to Mathematical Practices such as Justifying and Proving and Collaborative Mathematics, where success rates were lowest and rejection rates highest.

The results also highlight the importance of LLM selection. The transition from GPT-4 to GPT-5.2 resulted in substantial improvements in item quality, suggesting that advances in model capability can meaningfully improve item generation, particularly for more complex tasks. However, even with these improvements, all items required human revision, indicating that current models do not yet fully capture the requirements of high-quality mathematics assessment design without a human in the loop.

Importantly, this study contributes to the literature by examining AI-generated item development in a non-template-driven context. Much of the prior research on automated item generation has relied on structured or model-based approaches, where predefined templates enable control over item features such as difficulty and cognitive demand (Sonnleitner et al., 2025; Tan et al., 2025). In contrast, NAEP mathematics item development requires each item to be uniquely constructed to assess a specific intersection of characteristics, including content, practice, and item type. The Wilbur evaluation demonstrates that while LLMs can generate viable content-aligned items in this more flexible design space, they encounter greater difficulty when required to independently satisfy complex, multidimensional measurement targets without the scaffolding provided by templates.

These findings suggest that AI systems are best positioned as tools for supporting early-stage item generation and ideation, rather than as replacements for expert item writers. This aligns with broader conclusions in the literature that AI is most effective when integrated into human-in-the-loop development processes, where expert review is essential for ensuring construct alignment, correctness, and validity. In the context of NAEP mathematics, these needs are amplified by the

requirement to assess Mathematical Practices, which involve reasoning, explanation, and the integration of multiple aspects of mathematical thinking.

6.1 Future Evaluations

For future evaluations of Wilbur, prompts will be refined using secure mathematics content aligned with the 2026 NAEP Mathematics Framework, which will provide stronger exemplars for targeting both content objectives and NAEP Mathematical Practices. Building on the present findings, further work will also examine Wilbur's ability to generate selected-response formats other than SSMC, as well as multi-part, technology-enhanced, and constructed-response items. Continued testing with newer models will be especially important, given the gains observed when moving from GPT-4 to GPT-5.2 and the likelihood that improved instruction-following and coherence may further strengthen early draft generation. Thus, future Wilbur evaluations should move beyond asking whether generative AI can support assessment development to identifying the conditions under which it can do so efficiently, responsibly, and with sufficient quality to meaningfully support expert item writers.

7 Conclusion

The findings indicate that Wilbur is most effective as a tool for generating content-aligned draft items, while its performance declines when tasks require alignment to Mathematical Practices. These items are more difficult for AI because they must do more than state correct mathematics: they must elicit reasoning, support interpretation, and align multiple features of the item, such as the prompt, response options, and intended construct, within a single coherent task. This challenge was especially evident for practices such as Justifying and Proving and Collaborative Mathematics, which depend on nuanced reasoning and more complex construct alignment than content-focused items alone. The transition from GPT-4 to GPT-5.2 suggests that newer models may better support these demands by producing more coherent and instruction-responsive drafts, although expert review remains essential. As model capabilities continue to improve, AI systems may become more useful for generating early versions of practice-aligned items, but reliable alignment with complex mathematical

constructs will likely continue to require human oversight.

The improvement in results after the transition to GPT-5.2 demonstrates the critical role that model selection plays in AI-assisted item generation. Despite these improvements, all items continued to require revision, reinforcing the importance of expert oversight.

These results suggest that AI systems should be used strategically in assessment workflows, particularly for early-stage drafting, while more complex construct alignment remains dependent on human expertise.

Future work should focus on improving the extent to which AI systems can support the generation of practice-aligned items. Enhanced model training using secure, framework-aligned exemplar items, as well as the development of more sophisticated LLMs, may help accomplish this goal. Additional research is also needed to explore hybrid approaches that combine the flexibility of LLM-based generation with the structure of template-based approaches, particularly for supporting the development of items that require complex reasoning and diagnostic precision.

In conclusion, the Wilbur study provides evidence that AI can support mathematics item development within a complex assessment framework but also highlights the current limitations of these systems in capturing the processes central to mathematical proficiency. By examining AI performance within a non-template-driven context, this work advances our understanding of both the capabilities and constraints of generative AI in assessment development and identifies key directions for future research and implementation.

8 Limitations

Several design constraints should be considered when interpreting the results of this initial phase of AI item generation and evaluation. The study was intentionally limited in scope. Item generation focused on SSMC items, covered a restricted set of measurement targets, did not examine item-set generation, and included only a limited exploration of multi-part items. To keep the development process within practical limits, item refinement was restricted to three rounds of structured feedback.

The materials available for prompt development also shaped the results of the study. Secure mathematics items that could have supported more precise targeting of the 2026 NAEP Mathematics Framework could not be used, so framework-aligned exemplars were not incorporated to refine prompts, specifically to illustrate how items are used to assess the NAEP Mathematical Practices.

Furthermore, this study was structured to assess the usability of AI-generated drafts through expert review rather than through empirical student-response data. As a result, the findings provide evidence about draft quality under bounded development conditions, but not about student performance, the empirical behavior of distractors, or the validity of score interpretations in operational settings.

References

- ETS Assessment Content Development. (2026). Wilbur Evaluation Report: 2026 Reading Grades 4, 8, and 12; 2026 Mathematics Grades 4 and 8.
- Meissner, T., Schmidt, F., & Seifried, J. (2024). Large language models for competence-based item generation in mathematics education: Development and evaluation of Item Forge. *Frontiers in Education*, 9, Article 1324567. <https://doi.org/10.3389/feduc.2024.1324567>
- National Assessment Governing Board. (2021). Mathematics framework for the 2026 National Assessment of Educational Progress. U.S. Department of Education.
- Shah, R., Patel, A., & Kumar, S. (2025). Generating difficult mathematics questions with large language models through skill composition. *Proceedings of the Association for Computational Linguistics (ACL)*.
- Sonnleitner, P., Bernard, S., Michels, M. A., Inostroza-Fernandez, P., Keller, U., Gierl, M. J., Cardoso-Leite, P., & Hornung, C. (2025). Establishing cognitive item models for fair and theory-grounded automatic item generation: A large-scale assessment study with image-based math items. *Applied Measurement in Education*, 38(2), 95–117. <https://doi.org/10.1080/08957347.2025.2563889>
- Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. J. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2), 317–340. <https://doi.org/10.21449/ijate.1602294>