

Evaluating an AI-Item Generation Tool for NAEP Reading

Cara DiMeglio

ETS Research Institute, Princeton, NJ

[cdimeglio@ets.org]

Abstract

This study evaluated use of an AI-enabled item-generation tool for Reading item development for the National Assessment of Educational Progress (NAEP). Results were strongest when items targeted textually explicit information and weakest when passage and construct complexity increased. The tool showed potential for early-stage drafting rather than autonomous item development.

1 Background

Recent research suggests that generative AI may be useful for early-stage item development, but its role in educational assessment remains dependent upon diligent human oversight. Reviews of large language models (LLMs) show clear potential for generating candidate items at large scale, while also highlighting lasting concerns about validity, fairness, reliability, and construct representation in operational settings (Arslan et al., 2024; Tan et al., 2025). These concerns are especially significant for reading assessments, where effective items must be closely anchored to the text and accurately measure the intended construct. Evidence from reading-related studies further suggests that while LLMs can produce usable draft reading items, quality varies by task complexity, question type, and review conditions (Perkoff et al., 2023). Across this literature, the clearest point of consensus is the continued need for human judgment as AI-generated assessment content may improve efficiency and broaden the pool of candidate items, but expert review remains necessary to ensure overall item quality. This emerging evidence provides a strong rationale for the present evaluation of generative AI as a support tool for item development rather than as a substitute for expert item writers.

The 2026 Reading Framework for the National Assessment of Educational Progress positions NAEP Reading as a comprehension assessment organized around three disciplinary contexts—literature, science, and social studies—and four comprehension targets. These comprehension targets include Locate and Recall, which focuses on retrieving explicitly stated information; Integrate and Interpret, which requires readers to connect ideas and draw meaning across one or more texts; Analyze and Evaluate, which requires readers to examine textual quality, claims, craft, or evidence; and Use and Apply, which asks readers to use what they have understood from text for a broader purpose or task. Within each comprehension target, NAEP Reading also measures knowledge of Language Structures and Vocabulary, which focuses on the meaning and function of words, phrases, and language features in context (National Assessment Governing Board, 2021).

In practice, NAEP Reading item development begins with passage selection, after which Reading content developers establish the purpose for reading, map the intended testing points, and draft items aligned to framework and item specifications. Draft items then move through an iterative development cycle that typically includes content, fairness, editorial, committee, and client reviews, with revisions made at each stage until the items are suitable for further development or operational use. Because this process is labor-intensive and early drafting often requires generating multiple candidate items before expert refinement, the NAEP Reading team chose to explore whether Wilbur, an AI-enabled framework built by ETS and designed to scale and enhance stages of the assessment development lifecycle, could accelerate early-stage drafting, expand the pool of candidate items, and support more efficient development. Wilbur leverages multiple LLMs to generate and refine tailored content, aligned to each

program's custom requirements. The NAEP Reading team sought to explore whether they could utilize Wilbur's family of GPT models to produce usable draft items while still preserving the expert judgment required to ensure all items are aligned to NAEP's rigorous quality standards.

2 Key Questions

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

1. To what extent can Wilbur generate reading items that reliably meet NAEP expectations for quality, validity, and fairness?
2. How does performance vary by grade level, prompt structure, and construct?
3. Which features of the NAEP Reading assessment create challenges for AI-enabled item generation?

In this study, quality was defined narrowly and deliberately as expert-judged draft usability against a pre-established rubric. NAEP Reading team members evaluated whether each Wilbur-generated item represented a valid testing point (the content being tested was relevant, accurate, and essential for developing understanding of the passage), contained an identifiable key (a secure, defensible single best answer), and aligned with targeted constructs (as outlined in the 2026 Reading framework). Additionally, all items needed to meet program-defined expectations for clarity, fairness, accessibility, and style. The study did not collect student-response data and therefore does not claim evidence about operational measurement properties such as difficulty, discrimination, or validity based on student performance. Instead, the objective was to determine whether Wilbur could produce usable draft items that merited further human development within NAEP Reading's established content-review workflows.

3 Samples

The sample for this evaluation consisted of 450 Wilbur-generated NAEP Reading items across grades 4, 8, and 12, with 150 items at each grade. Item formats were limited to single-select multiple-choice (SSMC) and multiple-select multiple-choice (MSMC) to keep the scope manageable and to align with selected-response formats commonly used in operational NAEP Reading sets. The study focused on three 2026 Reading framework targets, Locate and Recall, Integrate and Interpret, and Language Structures and Vocabulary, which together provide a meaningful range of construct complexity while remaining feasible for an initial evaluation of AI-enabled discrete item generation. At each grade level, 60 Locate and Recall items, 60 Integrate and Interpret items, and 30 Language Structures and Vocabulary items were generated.

Item generation was grounded in twenty-four released NAEP Reading passages distributed relatively evenly across literary, science, and social studies disciplinary contexts. Because available released materials had been coded to an earlier assessment framework¹, passages were recoded for the 2026 Reading framework before use. To maintain control over the evaluation, the study was limited to discrete items tied to single passages rather than full reading blocks or scenario-based tasks. The more complex framework targets, Analyze and Evaluate and Use and Apply, as well as more complex item structures, were excluded from this initial evaluation. This was largely due to time constraints but also a desire to prioritize straightforward item types before addressing more complex formats, enabling a more manageable analysis of Wilbur's capabilities.

4 Methods

4.1 Item Generation Procedures

Wilbur-generated draft items in batches of ten items for each passage using GPT-4, which was the default model at the start of the evaluation. For consistency in evaluation conditions, the NAEP Reading team retained GPT-4 throughout the study, even after more advanced models became available.

¹ The NAEP Reading framework specifies the reading knowledge and skills to be measured at grades 4, 8 and 12 and guides item development, scoring, and reporting across these grades. The 2026 framework was updated from the

2009 iteration to reflect advances in reading research, the implementation of digitally based testing, and contemporary literacy demands.

4.2 Prompt Conditions

Two prompt conditions were implemented: constrained and free rein. The constrained prompt required Wilbur to generate five SSMC and five MSMC items and to distribute them across specified subskills derived from the 2026 Reading framework; additional restrictions were included to reduce content overlap and to ensure coverage of at least one main idea item in each set. By contrast, the free rein prompt imposed relatively few restrictions. Wilbur was asked to generate ten SSMC or MSMC items aligned to the selected comprehension target, without prescribed distributions of item types or subskills. The comparison between these conditions was designed to examine how stricter versus more flexible prompt parameters influenced item quality.

4.3 Iterative Refinement

Generated items were revised through a maximum of three structured feedback loops within the Wilbur interface. Early feedback focused on establishing valid testing points and defensible keys, and alignment with the intended construct. Later rounds addressed distractor quality, clarity of stems and options, and adherence to stylistic, fairness, and accessibility expectations. This structure allowed the study to assess not only initial Wilbur output but also the extent to which iterative feedback could improve draft usability.

4.4 Evaluation Procedures

All items were evaluated with a two-tier rubric consisting of three essential criteria and five secondary dimensions (Table 1). The rubric was adapted from established NAEP item review criteria and served as the study's expert-review instrument. The essential criteria, a valid testing point, an identifiable key, and alignment with targeted constructs, defined the minimum threshold for draft usability, while the secondary dimensions identified revisable weaknesses in wording, option quality, grade appropriateness, fairness, and NAEP style. Items were scored on a 15-point scale with a cut score of 10 points. Each essential criterion was worth 1 point and each secondary criterion was worth 3 points. Items that scored 12 or above were considered acceptable with minor revisions while items that scored 10-11 points were considered acceptable with major revisions. Items scoring 9 points or less were rejected as requiring too much revision to be viable or efficient for use. Items were

considered acceptable for further development only when reviewers judged the essential criteria to be met and any remaining weaknesses to be revisable within standard development practice.

Type	Criterion
Essential 1	Valid testing point
Essential 2	Identifiable key
Essential 3	Construct alignment
Secondary	Stem clarity
Secondary	Key and distractor quality
Secondary	Grade appropriateness
Secondary	Fairness/sensitivity
Secondary	NAEP style adherence

Table 1: Rubric criteria by type

Items were reviewed by subject matter experts using the rubric and a structured consensus process. All items were double-scored, with some triple-scored to support reviewer alignment and consistency. The analyses reported here are based on the final consensus determinations and should therefore be interpreted as expert-review outcomes within a content-evaluation framework.

5 Results

5.1 Overall Results

Across all conditions, 305 of 450 reading items (68%) were judged acceptable for further development. These items were accepted with varying levels of revisions needed, and no items were assessment-ready without further human intervention.

5.2 Results by Grade

By grade, acceptance rates declined from 73% at grade 4 to 68% at grade 8 and 62% at grade 12 (Table 2). This downward trajectory suggests that Wilbur had greater difficulty producing usable draft items as passage demands and construct complexity increased across grade levels.

Grade	Acceptance by Item Count	Acceptance by Percentage
4	110/150	73%
8	102/150	68%
12	93/150	62%

Table 2: Acceptance by grade level

5.3 Results by Prompt

By prompt condition, constrained and free rein prompting produced similar acceptance rates at grade 4 (73% vs. 74%, Cohen's $h = -0.02$), whereas constrained prompting produced higher acceptance rates at grade 8 (73% vs. 64%, Cohen's $h = 0.19$) and slightly higher at grade 12 (63% vs. 61%, Cohen's $h = 0.05$) (Table 3).

Although constrained prompting produced somewhat stronger, though negligible based on Cohen's h (Cohen, 1988), results at grades 8 and 12, these differences should be interpreted cautiously because the two prompt conditions were not designed as a controlled experimental comparison.

Grade	Prompt Type	Acceptance by item count	Acceptance by percentage
4	Constrained	51/70	73%
4	Free rein	59/80	74%
8	Constrained	44/60	73%
8	Free rein	58/90	64%
12	Constrained	38/60	63%
12	Free rein	55/90	61%

Table 3: Acceptance rates by prompt type

5.4 Results by Construct

By construct, Locate and Recall items had the highest acceptance rates overall, with strong results at grades 4 (92%) and 8 (93%) and a lower, though still comparatively strong, rate at grade 12 (77%) (Table 4).

Grade	Acceptance by Item Count	Acceptance by Percentage
4	55/60	92%
8	56/60	93%
12	46/60	77%
Overall	157/180	87%

Table 4: Acceptance of Locate and Recall items across grades

Integrate and Interpret items showed moderate acceptance rates across grades, ranging from 57%

to 72% (Table 5), while Language Structures and Vocabulary items had the lowest acceptance rates overall, ranging from 37% to 53% (Table 6).

Grade	Acceptance by Item Count	Acceptance by Percentage
4	43/60	72%
8	34/60	57%
12	40/60	67%
Overall	117/180	65%

Table 5: Acceptance of Integrate and Interpret items across grades

Grade	Acceptance by Item Count	Acceptance by Percentage
4	12/30	40%
8	11/30	37%
12	16/30	53%
Overall	39/90	43%

Table 6: Acceptance of Language Structures and Vocabulary items across grades.

5.5 Reviewer Agreement

Reviewer agreement was examined using percent agreement on initial item-level acceptance decisions (Accept/Reject) made during the review process. All 450 items were reviewed by at least two subject-matter experts, and some items were triple-scored to support reviewer calibration and alignment. Because the analyses reported here focus on percent agreement rather than a chance-corrected reliability coefficient, the findings should be interpreted as descriptive evidence of reviewer consistency within this expert-review process.

Percent agreement ranged from 73% to 80% across grade levels, with an overall agreement rate of 77% (Table 7). Grade 4 demonstrated the highest agreement rate (80%), while grade 12 showed the lowest (73%), suggesting that items at higher grade levels may present greater evaluative complexity for reviewers.

Grade	Agreement by Item Count	Agreement by Percentage
4	120/150	80%
8	117/150	78%
12	110/150	73%
Overall	347/450	77%

Table 7: Reviewer agreement by grade level

Constrained items showed somewhat higher agreement than free rein items overall (80% vs. 75%), as shown in Table 8. This pattern suggests that more structured prompting may have yielded outputs that were easier for reviewers to evaluate consistently, although this comparison should be interpreted cautiously given the broader design differences between prompt conditions.

Prompt Type	Grade 4	Grade 8	Grade 12	Overall
Constrained	79%	80%	81%	80%
Free rein	80%	77%	68%	75%

Table 8: Reviewer agreement by prompt type

Agreement also varied by construct (Table 9). Locate and Recall items showed the highest reviewer agreement with an overall agreement rate of 91%, followed by Integrate and Interpret items (77%). Language Structures and Vocabulary items showed the lowest agreement (53%), suggesting that these items were more difficult to evaluate consistently and may require clearer criteria, additional reviewer calibration, or both.

Construct	Grade 4	Grade 8	Grade 12	Overall
L&R	92%	93%	89%	91%
I&I	78%	75%	77%	77%
LSV	57%	53%	50%	53%

Table 9: Reviewer agreement by construct

Overall, the reviewer agreement results suggest reasonably consistent expert judgments within the study's content-review framework, while also identifying areas of greater evaluative difficulty. Agreement was lower for higher-grade items, lower for items generated using the free rein prompts, and lowest for Language Structures and Vocabulary items. These patterns reinforce the broader results of the study by suggesting that the most challenging item types, those based on the

most complex constructs or linked with more complex passages, were difficult not only for Wilbur to generate well, but also for reviewers to evaluate consistently.

6 Discussion

The evaluation results point to both meaningful strengths and clear limitations in Wilbur's current performance for NAEP Reading. NAEP Reading Team members noted that Wilbur was largely successful in generating Locate and Recall items, particularly at grades 4 and 8, and that it produced SSMC and MSMC items with moderate success overall. Many generated items met basic authoring expectations, including a defensible correct answer and the expected structure for the intended item type. Because multiple-choice items make up a substantial proportion of NAEP Reading, these findings suggest that Wilbur may have practical value for early-stage ideation and draft generation of selected-response items within a human-centered development process. Reviewers also observed that Wilbur occasionally generated items that provided a sophisticated or unexpected approach to a comprehension target, such as focusing on a subskill less typically used (under free rein prompt conditions). For example, a Locate and Recall item asking students to "Infer the meanings of unfamiliar words using information in the sentences immediately surrounding the word" or an Integrate and Interpret item asking students to "describe how the setting of the story contributes to the theme," offered opportunities for enhanced variety in how these constructs are measured.

At the same time, the results indicate that Wilbur was less successful as construct demands increased for each grade level and as passage difficulty increased with grade level expectations. Although Locate and Recall items remained recognizable and largely usable across grades, results were weaker at grade 12, and coverage of the full range of skills within that target was limited. Integrate and Interpret and Language Structures and Vocabulary items proved more challenging, with many generated items lacking sufficient substance, misinterpreting passage information, or failing to reflect the breadth and depth of the intended 2026 framework targets. Reviewers also noted that Wilbur sometimes overfocused on a narrow aspect of a passage or testing point, leading to repetitive item sets and less productive feedback loops. These weaknesses likely reflect, at least in part, the

limited availability of framework-aligned training materials as well as the increased complexity resulting from the heavy reliance on passage text for item generation, which is a requirement more characteristic of NAEP Reading than of other NAEP subjects.

The Language Structures and Vocabulary results were especially weak across grades, frequently failing to meet essential criteria such as passage dependency. In addition, after grade 4 Language Structures and Vocabulary items showed particularly poor performance using the constrained prompt, later grade 8 and 12 item generation was shifted to the free rein prompting conditions to give the system more opportunity for success with this demanding construct.

Taken together, the findings suggest that Wilbur's performance may improve with additional training and access to more representative framework-aligned materials, but that further work will be needed before it can support more complex reading structures, including reading blocks developed around a predefined purpose for reading and multi-passage sets.

6.1 Implications for Future Evaluations

Future evaluations of Wilbur will focus on extending the scope of the current study to better determine how the system can support larger-scale NAEP item development for future development cycles. In Reading, this includes moving beyond discrete multiple-choice items tied to single passages and examining whether Wilbur can effectively support more complex reading structures, such as full reading blocks and scenario-based tasks that require a coherent purpose for reading and cross-textual items.

Future work will also explore how newly granted access to more representative framework-aligned materials for training purposes, continued prompt refinement, and updated model selection can improve performance on more complex constructs and item types. Taken together, these next steps position future Wilbur evaluations as an effort to define not only whether and how generative AI can increase efficiency, but also the conditions under which it can responsibly support higher-quality assessment development.

7 Conclusions

This study examined the feasibility of AI-enabled item generation in NAEP Reading through a structured, human-led evaluation of Wilbur, an ETS-developed AI framework. The findings show moderate success in producing usable draft multiple-choice items, particularly for lower-complexity targets, but they also demonstrate clear limits to AI-enabled item generation, as item quality declined as construct complexity and grade-level reading demands increased. Reading remains a complex context for AI-enabled item generation because quality depends not only on producing plausible, construct-aligned assessment questions, but also on grounding those questions securely in the source text(s).

As such, the key takeaway of the study is that generative AI models (facilitated by Wilbur) appear most useful as a support tool for early-stage drafting and ideation of NAEP Reading items within a human-centered development workflow. Generative AI can supplement the development of NAEP Reading items, but outputs will still require expert human review and revision. Wilbur's weaknesses were most evident when constructs were more complex, indicating that continued training, prompt refinement, and updated model selections will be needed to improve outputs for all items aligned to the 2026 Reading framework. These findings position Wilbur, and generative AI more broadly, not as a substitute for Reading content specialists, but as a tool whose value depends on careful prompting, constrained use, and sustained human judgment.

8 Limitations

The findings should be interpreted within the design constraints of the Reading evaluation. Most importantly, contrasts between constrained and free rein prompting should be understood as documented design conditions rather than as clean experimental manipulations, particularly because prompting conditions shifted for some Language Structures and Vocabulary items as the study proceeded in an attempt to increase the quality of item output. The prompt comparison should likewise be interpreted cautiously, because the constrained and free rein prompt conditions differed not only in the level of specificity of the prompts but also in the degree to which NAEP

Reading team members pre-selected the subskills and quantities for a particular set of items.

More broadly, the evaluation was limited to discrete multiple-choice items based on single passages rather than full reading blocks or scenario-based tasks. As a result, it did not assess Wilbur's capabilities in generating items within larger NAEP reading structures that depend on an established purpose for reading, a reader role, and/or coordinated item development across a block.

The source materials also imposed important constraints on the results of the evaluation and subsequent interpretations. Due to the secure nature of the NAEP assessment, this initial study had to rely on released NAEP passages that were not aligned with the 2026 Reading framework. Because the associated released items were also not aligned to the 2026 framework, they could not serve as framework-aligned exemplars, limiting the ability to robustly train Wilbur prior to the start of item generation.

The evaluation was also limited by design to expert review of generated drafts for usability only. Revisions were capped at three structured feedback loops to reflect a practical threshold for development efficiency. The findings therefore support conclusions about content quality and draft usability under deliberately constrained development conditions; they do not establish how the items would function with students, whether distractors would perform as intended, or whether accepted items would yield valid score interpretations in an operational setting. These constraints were purposeful and should be read as a clear statement of evidentiary scope rather than as a deficiency of the study.

Finally, the generation of items was conducted using GPT-4, the default model at the start of the evaluations. During and since the completion of the evaluation, more advanced LLMs have become available for use, many with deeper analytical thinking capabilities better aligned with complex item generation tasks. Thus, it can be anticipated that some of the earlier deficiencies in Wilbur-produced items will be remedied by use of newer LLMs for future item generation. As such, the quality of items produced during the initial Wilbur evaluation should be viewed as a baseline rather than a definitive reflection on AI's capability to generate quality draft items for operational use.

References

- Arslan, B., Lehman, B. A., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence*, 7, Article 1460651. <https://doi.org/10.3389/frai.2024.1460651>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- ETS Assessment Content Development. (2026). *Wilbur Evaluation Report: 2026 Reading Grades 4, 8, and 12; 2026 Mathematics Grades 4 and 8*. [Unpublished internal report]
- National Assessment Governing Board. (2021). Reading framework for the 2026 National Assessment of Educational Progress. U.S. Department of Education. [Reading Framework for the 2026 National Assessment of Educational Progress](#)
- Perkoff, E. M., Bhattacharyya, A., Cai, J. Z., & Cao, J. (2023). Comparing neural question generation architectures for reading comprehension. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 556–566). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bea-1.47>
- 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>