

Evaluating an AI-Item Generation Tool for NAEP Science

Carolina Safar, Mitchell Price, and Jeffery Ackley

ETS Research Institute, Princeton, NJ

[csafar@ets.org, mprice001@ets.org, jackley@ets.org]

Abstract

AI-enabled systems could transform item authoring for standardized assessments. We compare framework alignment and accuracy of NAEP Science items developed by two AI tools and human authors. Items from all sources had similar acceptance rates and most accepted items would require significant revision, underscoring the importance of evaluating AI-generated content.

1 Background

The National Assessment of Educational Progress (NAEP) in Science has provided data about student achievement in science in the United States since 1969. The 2029 NAEP Science assessment will be the first using a multidimensional framework that elevates phenomena and engineering problems in the measurement. The 2028 NAEP Science Assessment Framework (“the NAEP Science framework”) was developed to better reflect current science education trends in the U.S. by emphasizing students’ ability to apply sensemaking to authentic contexts and phenomena, using disciplinary concepts (DCs), science and engineering practices (SEPs), and crosscutting concepts (CCCs).

NAEP Science item development has historically been a human-centered, iterative process. The first round of item development for the new NAEP Science framework included items that were initially authored by external subject matter experts with extensive experience in science education and assessment, hereinafter referred to as outside item writers (OIWs). These initial items are revised in conjunction with NAEP Science assessment specialists. Items that are accepted for further development are subjected to multiple iterative internal reviews, external expert-panel reviews, as well as pilot testing and analysis before

items are considered for the operational assessment.

Recent published research on generative AI in assessment development suggests that large language models may help reduce the time required to draft assessment items, but that their outputs still require domain-specific review for validity, accuracy, and framework alignment. (Tan et al., 2025). Nyaaba, Zhai, and Faison (2024) explored generative AI for K–12 science assessment by developing a culturally responsive science assessment generator aligned to performance expectations from the Next Generation Science Standards (NGSS). Their pilot demonstration suggested that a customized GPT could generate culturally and contextually tailored science assessment items but also highlighted the need for expert review of cultural, contextual, and scientific validity.

In addition to formal research, publicly available AI tools have begun to target generation of multidimensional assessment items aligned to science standards. Other publicly available tools, such as Flint’s NGSS-aligned science assessment generator and Musely’s 3D Science Assessment Generator, describe functionality for producing assessments that integrate content and practice domains. These tools indicate growing interest in AI-supported three-dimensional science assessment development, but the available public descriptions generally do not provide evidence that generated items meet the level of framework alignment, scientific accuracy, accessibility, and psychometric quality required for large-scale assessment programs.

This report provides a detailed evaluation of assessment items targeting NAEP Science assessment criteria that were generated by two AI-enabled item generation tools: a proprietary ETS tool, Wilbur, and a publicly available NGSS Item Authoring Tool. The quality and usability of items

generated by these tools is compared with that of items generated by human OIWs.

2 Key Questions

The evaluation addressed four key components in assessing Wilbur’s capabilities to generate NAEP Science items, as follows:

1. To what degree do Wilbur-generated items meet the requirements of the 2028 NAEP Science Assessment Framework?
2. To what degree can Wilbur consistently generate items that contain accurate science content?
3. How does using Wilbur to generate items compare to using outside item writers (OIWs) in terms of item quality and efficiency?
4. How does using Wilbur to generate items compare to using a framework-focused AI tool, the NGSS Item Authoring Tool, in terms of item quality and efficiency?

3 Samples

3.1 Items Generated by Wilbur

The Wilbur evaluation included 144 generated discrete items across grades 4, 8, and 12, with 48 items generated and evaluated per grade. For each grade, the Science team selected two DCs from each of the three NAEP Science discipline groups: Physical Science, Life Science, and Earth and Space Sciences. The team also selected two SEP sub-statements per grade, one from the Analyzing category and one from the Explaining category.

The evaluation was limited to single-select multiple-choice (SSMC) and inline choice (ILC) item types. SSMC was selected because it represents the primary item type Wilbur was designed to generate and is commonly used in NAEP Science assessments. The ILC item type was selected because it is typically less complex to write items for and represents a meaningful item type within the NAEP Science pool.

The Science team intentionally focused on comparatively less challenging item-generation targets when selecting the targeted DCs, SEPs, and item types. Crosscutting concepts, the third dimension in the NAEP Science framework, and

item sets (multiple items linked to one shared stimulus) were excluded so that the evaluation could focus on two-dimensional discrete items and reduce prompt complexity.

3.2 Items Authored by Outside Item Writers

For comparison, the Science team evaluated 11 OIW-authored items: six grade 4 items and five grade 8 items. Grade 12 was excluded because grade 12 item development for the 2028 NAEP Science framework had not yet occurred. Prior to authoring the items, OIWs received extensive NAEP-specific training. The selected OIW-authored items were drawn from a recent grade 4 and grade 8 OIW development cycle and were restricted, as much as possible, to items with the same or similar metadata as the targeted Wilbur metadata. While this restriction did limit the pool of OIW-authored items available for analysis, it supported a more appropriate comparison between these items and Wilbur-generated items.

3.3 Items Generated by the NGSS Item Authoring Tool

The NGSS Item Authoring Tool was included as a comparison because it is a customized generative AI application developed to support the creation of three-dimensional assessment items aligned to the NGSS (Reiter, 2025). Like the 2028 NAEP Science framework, the NGSS framework is a set of K-12 science standards grounded in *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* from the National Research Council/National Academies.

The Science team evaluated 16 grade 8 items generated using the NGSS Item Authoring Tool. The team targeted one of the least and one of the most challenging grade 8 DCs for Wilbur for this comparison, both of which were in the Earth and Space Sciences discipline. These two DCs were targeted with the same grade 8 SEP and item type combinations used in the Wilbur evaluation.

4 Methods

4.1 Prompt Development and Item Generation

The prompts used for item generation were developed in collaboration between the Science team and Wilbur prompt engineers. The Science team drafted initial prompts that emphasized critical components of both the NAEP Science

framework requirements and assessment best practices, and Wilbur prompt engineers used provided framework resources for prompt refinement. The prompts were tested and further refined by implementing feedback from the Wilbur prompt engineers based on early Wilbur output and interactions. During prompt testing and refinement, GPT-5.2 was made available. The Science team experimented with GPT-4 and GPT-5.2 and ultimately decided to use GPT-5.2 for this study due to its increased capabilities for writing context-based items that leverage scientific data and integrate scientific reasoning.

Due to the secure nature of operational NAEP Science items, this initial study did not use secure framework-aligned exemplar items as training material for Wilbur. An additional challenge was identifying an effective balance in the length and complexity of prompts and feedback; providing sufficient detail to support focus on NAEP-specific requirements while maintaining enough brevity to avoid diminishing returns in item quality. Wilbur prompt engineers guided the NAEP Science team members in refining and focusing the prompts as was possible.

Two final prompts were developed, one for SSMC items and one for ILC items. Both prompts included placeholders for grade level, DC statement, and SEP sub-statement. The final prompts were used with Wilbur's GPT-5.2 model throughout the formal evaluation.

For each item-generation session, Wilbur was prompted to generate two items for a specified item type, DC, and SEP combination. Reviewers could provide targeted feedback to Wilbur for up to three feedback loops. After iterative feedback loops were completed, reviewers selected the version of each item best aligned with the intended assessment targets and evaluated that version using the study rubric. The NAEP Science team chose to evaluate the best version of each item rather than the final version to both evaluate the highest demonstrated capabilities of the tool and to mirror the practice that would be enacted with human-authored versions of an item.

The NGSS Item Authoring Tool was given the same prompts for item generation and the same level of feedback for iterations: two items were generated at a time, feedback loops were limited to a maximum of three, and the highest-quality version of each item was selected for evaluation.

4.2 Evaluation Rubric

Items from all three sources were evaluated using the same standardized rubric developed for NAEP Science SSMC and ILC items. The rubric included nine criteria grouped into two categories: 2028 NAEP Science Framework-specific Requirements and Assessment Best Practices. The Framework-specific criteria were inclusion of context and phenomenon, science content accuracy, DC alignment, and SEP alignment. The Assessment Best Practice criteria were grade-level content and language appropriateness, stem clarity and appropriateness, key accuracy and distractor/rationale quality, absence of bias or sensitivity concerns, and adherence to NAEP editorial and style guidelines.

Framework-specific criteria were weighted more heavily than Assessment Best Practice criteria to better align the evaluation with the core requirements unique to the NAEP Science framework. Items received total scores from 13 to 39 points, and a score above 25 points (the midpoint of the rubric effective range) was set as the threshold for acceptance. Items scoring 25 points or lower were rejected.

In addition to receiving rubric scores, items were assigned readiness ratings to indicate how ready they would be for further use in the established NAEP Science item development pipeline. A rating of 3 indicated that the item was usable as is, 2 indicated usable with minor revisions, 1 indicated usable with major revisions, and 0 indicated unusable due to cumulative low scores or the presence of a critical flaw. Readiness was evaluated separately from rubric scores because the effort needed for revision depends on which criteria received low scores and how readily those issues can be remedied. All items were double-scored or triple-scored. A third review was performed for items with the greatest differences between initial scores. Final scores and ratings were calculated by averaging each item's individual reviewer scores and ratings, except for items with three reviews, which were evaluated by the group to determine the most appropriate final scores and ratings.

5 Results

5.1 Wilbur-Generated Item Acceptance and Readiness

Of the 144 total items generated by Wilbur, 63 items (44%) were accepted while 81 items (56%) were rejected. The majority of accepted items were rated as requiring major revisions, and none were rated usable as is. Acceptance rates were similar by grade, ranging from 40% to 50% with the highest acceptance rate in grade 12. Table 1 shows the acceptance and rejection rates for Wilbur-generated items.

Grade	Accepted	Rejected	Total n
4	20 (42%)	28 (58%)	48
8	19 (40%)	29 (60%)	48
12	24 (50%)	24 (50%)	48
Total	63 (44%)	81 (56%)	144

Table 1: Wilbur-generated item acceptance and rejection rates by grade

Table 2 shows the readiness ratings of Wilbur-generated items by grade.

Grade	Usable As Is	Usable with Minor Revisions	Usable with Major Revisions	Rejected
4	0 (0%)	5 (10%)	15 (32%)	28 (58%)
8	0 (0%)	4 (8%)	15 (32%)	29 (60%)
12	0 (0%)	6 (12%)	18 (38%)	24 (50%)
Total	0 (0%)	15 (10%)	48 (34%)	81 (56%)

Table 2: Wilbur-generated item readiness by grade

5.2 Results by DC, SEP, and Item Type

Wilbur-generated item acceptance rates varied most by DC when compared with acceptance rates by other targeted metadata. By DC, acceptance rates ranged from 12% for E4.7, E8.6, and E12.6 to 87% for P4.3. In each grade, the lowest acceptance rate occurred for the DC involving water cycling, weathering, and erosion. Wilbur-generated items for this subtopic often lacked sufficient alignment with the concept and conflated content from Physical Science DCs about phase changes.

By SEP category, differences were smaller. Across grades, 46% of items targeting the Analyzing practice category were accepted, compared with 42% of items targeting the Explaining practice category. This pattern suggests that targeted concept may have had a greater impact on Wilbur-generated item quality than targeted practice, although only two SEP categories were included in the evaluation and therefore interpretation of the results is limited.

Item type also affected acceptance rates. Across grades, 50% of SSMC items were accepted, compared with 38% of ILC items. This pattern may reflect Wilbur's primary design as an SSMC authoring agent, with ILC-specific instructions supplied through prompts rather than through preexisting internal programming.

5.3 Multidimensionality and Framework-Specific Scores

Among Wilbur-generated items, fewer than one-third successfully integrated the targeted DC and SEP. Specifically, 29% of the Wilbur-generated items were accepted and met the study's criterion for being minimally two-dimensional. Items were identified as minimally two-dimensional if they received an average score of at least 4 out of 6 points for both the DC Alignment and the SEP Alignment rubric categories. Table 3 shows the dimensionality of accepted Wilbur-generated items.

Grade	2D (DC & SEP)	1D (DC or SEP)
4	14 (29%)	6 (13%)
8	15 (29%)	4 (8%)
12	13 (27%)	11 (23%)
Total	42 (29%)	21 (14%)

Table 3: Dimensionality of accepted Wilbur-generated items

Average categorical scores for accepted Wilbur-generated items were near the midpoint of the 2-to-6-point framework-specific scale. Across grades, the average score was 4.14 for Context and Phenomenon, 4.23 for Science Content, 3.95 for DC Alignment, and 3.88 for SEP Alignment. These scores reinforce the finding that true multidimensional alignment was a central challenge.

5.4 Comparison with OIW's and the NGSS Item Authoring Tool – Item Quality

The OIW component of the evaluation included 11 items. Of these, five items, or 45%, were accepted, and six items, or 55%, were rejected. Acceptance rates were 50% for grade 4 OIW-authored items and 40% for grade 8 OIW-authored items, similar to the range observed for Wilbur-generated items.

OIW-authored and Wilbur-generated items received comparable categorical scores for the framework-specific parameters. However, OIW-authored items showed greater diversity in overall item structure than Wilbur-generated items. Wilbur-generated items more often followed repetitive, template-like formats with similar question construction across items. Our experience with large-scale item development with OIW-authored items resulted in a wide range of item contexts and styles; however, the small OIW-authored item sample in this analysis limited finer-grain analyses of additional metadata.

The NGSS Item Authoring Tool generated 16 grade 8 items, of which 10 items, or 62%, met the minimum score for acceptance. Seven of eight E8.5 items were accepted, compared with three of eight E8.6 items. This pattern was similar to the Wilbur

pattern for these DCs, with stronger performance in targeting E8.5 than E8.6.

Of the three item authoring sources, the NGSS Item Authoring Tool had the highest overall acceptance rate, but its average categorical scores for Context and Phenomenon, Science Content, and DC Alignment were lower than Wilbur's. In contrast, the NGSS Item Authoring Tool had the highest average SEP Alignment score among the three sources. Table 4 shows acceptance and rejection rates by authoring source.

Source	Accepted	Rejected	Total n
Wilbur	63 (44%)	81 (56%)	144
OIW	5 (45%)	6 (55%)	11
NGSS Item Authoring Tool	10 (62%)	6 (38%)	16

Table 4: Acceptance and rejection rates by source

Table 5 shows the readiness ratings of items by authoring source.

Table 6 shows the average scores for the framework-specific rubric categories of accepted items by source.

Source	Usable As Is	Usable with Minor Revisions	Usable with Major Revisions	Rejected	Total n
Wilbur	0 (0%)	15 (10%)	48 (33%)	81 (56%)	144
OIW's	0 (0%)	1 (9%)	5 (45%)	5 (45%)	11
NGSS Item Authoring Tool	0 (0%)	3 (19%)	7 (44%)	6 (37%)	16

Table 5: Item readiness ratings by source

Source	Context & Phenomenon	Science Content	DC Alignment	SEP Alignment
Wilbur	4.14	4.23	3.95	3.88
OIW's	3.36	4.65	3.70	3.45
NGSS Item Authoring Tool	3.69	3.69	3.65	4.02

Table 6: Accepted item average categorical scores for framework-specific parameters by source

5.5 Comparison with OIW and the NGSS Item Authoring Tool – Efficiency

Wilbur was efficient at generating large numbers of draft items. Reviewers typically took between 30 and 45 minutes to generate, iterate, and evaluate each Wilbur-generated item. This metric is comparable to that of using the NGSS Item Authoring Tool for item generation, and significantly faster than using OIW for item authoring, which typically requires several days of lag time between a writing assignment and receiving the draft items.

At the same time, the effective yield of items ready for further development from any of the three sources was limited. Like items from the other authoring sources, Wilbur-generated items often required major revision, and no evaluated items were considered ready as is. Reviewers identified challenges with using Wilbur, including weak multidimensional integration, limited use of the original context or phenomenon in the item’s sensemaking, repetitive contexts and structures, weak distractors and rationales, and inefficient response to high-level feedback. Some of these challenges were also present with OIW and the NGSS Item Authoring Tool, while some were unique to Wilbur and other distinct challenges arose from using the other two item sources. The results suggest that Wilbur may be most useful for initial item generation, without internal revision, of low complexity items and potentially with some level of human-provided starting points to maximize known strengths and account for areas of weakness.

6 Discussion

6.1 Key Takeaways

Incorporating AI-enabled tools in content creation has the potential to alleviate some of the demands of developing large item pools for standardized assessments. This study takes the critical step of evaluating the quality of AI-assisted item development for a complex multidimensional science assessment framework. Wilbur showed promise as an early-draft NAEP Science assessment item generator, particularly for less complex item types and some DCs. However, the finding that only 29% of the full Wilbur pool consisted of accepted items meeting the study’s criterion for minimal two-dimensionality suggests

that multidimensional sensemaking remains a difficult target for AI item generation in science.

The comparison with OIW and the NGSS Item Authoring Tool contextualizes Wilbur’s performance. The overall acceptance rate for Wilbur-generated items was similar to that of OIW-authored items in this constrained comparison, while the acceptance rate for NGSS Item Authoring Tool-generated items was higher in a narrower grade 8 Earth and Space Sciences sample. However, all three sources produced items that would require significant revision before entering the first content-review phase of development.

For NAEP Science, the comparison across sources is especially informative because each source showed different strengths. OIW-authored items had the highest average Science Content score, consistent with the value of experienced science educators and assessment writers. Wilbur-generated items had higher average scores than NGSS Item Authoring Tool-generated items for Context and Phenomenon, Science Content, and DC Alignment, while NGSS Item Authoring Tool-generated items showed stronger average SEP Alignment.

This report illustrates the importance of evaluating AI-supported item generation tools against discipline-specific construct requirements rather than general item-writing quality alone. In NAEP Science, the central challenge is not simply generating plausible item text; it is generating items that authentically embed a phenomenon or engineering problem and require students to integrate disciplinary content with a science and engineering practice to make sense of the phenomenon or solve the problem. The use of AI-enabled tools for generating assessment content introduces both opportunities and responsibilities. While these tools have the power to produce large numbers of items in relatively short amounts of time, they require rigorous evaluation of the resulting content to ensure the myriad criteria for suitability are met.

6.2 Reflections on Methodology

This study represents the first detailed analysis of the scientific accuracy and framework alignments of AI-generated science assessment items. Every item was evaluated by a minimum of two independent raters using a standardized rubric with both NAEP Science framework-specific criteria and general assessment best practice criteria.

Reviewer scores agreed on the accept/reject classification for 101 of 171 items, or 59%. Across the full sample, reviewers' total scores differed by an average of 4.25 points on a 39-point rubric. Because the accept/reject classification depended on a fixed 25-point threshold, small score differences near the threshold may have contributed to classification disagreements. This is further explained by the small average difference between reviewer scores across all items and the small number of items that required a third review, which represented 8% of the full sample.

At the time of this evaluation, the grade 8 OIW-authored items included in the study had already progressed through a substantial portion of the item development life cycle and had reached the pilot stage. This provided an opportunity to compare each item's evaluation outcome—whether it met the rubric-based acceptance cutoff and, if accepted, the level of revision judged necessary—with the item's actual outcome in the NAEP Science development process. Overall, the evaluation results aligned with the real development outcomes for 80% of the OIW-authored items. In most cases, OIW-authored items that were accepted in the evaluation also advanced through the NAEP Science item development pipeline, and the level of revision identified during the evaluation generally corresponded to the extent of revision the items actually underwent.

6.3 Future Directions

Future evaluation phases should examine Wilbur's performance with secure framework-aligned exemplars, additional item types, item sets, constructed-response items, diagrams and visual components, and narrower co-authoring workflows in which Wilbur supports selected components of item development rather than full end-to-end item generation.

7 Conclusion

The NAEP Science evaluation suggests that Wilbur has potential as a support tool for early-stage NAEP Science item drafting and ideation, but it is not yet sufficient for full-scale item generation without substantial human review and revision. Continued internal Wilbur program training and prompt refinement is expected to support expanded use of Wilbur for item generation.

The evaluation contributes to AI-in-measurement research by foregrounding the

perspective of assessment specialists as expert users and evaluators of AI-generated content. It also demonstrates the importance of evaluating generative AI tools against the specific construct requirements of the target assessment.

8 Limitations

Several limitations constrain interpretation of the findings. The Wilbur evaluation used a limited subset of DCs, SEPs, and item types and intentionally focused on comparatively less challenging item-generation targets. Crosscutting concepts, item sets, constructed-response items, and graphic-generation features were not included.

The OIW and NGSS Item Authoring Tool comparison samples were smaller and more constrained than the Wilbur sample. While these design choices were made to balance the evaluation timeline with the focus on Wilbur, the reduced scopes do limit some levels of comparison between the three item writing sources.

Wilbur remained under active development at the time of the evaluation, and system limitations and instabilities may have affected both efficiency and output quality. Additionally, no information was available regarding the internal training of the NGSS Item Authoring Tool, thus limiting the ability to speak to the reason for any observed differences between it and Wilbur.

References

- National Assessment Governing Board. (2024). [2028 NAEP Science Assessment Framework](#). National Assessment Governing Board.
- Nyaaba, M., and Zhai, X. (2026). [Generative AI-based automated assessment generation](#). In *Artificial Intelligence for STEM Education Research: Advanced Methods and Applications*, pages 39–59. Springer Nature Switzerland.
- Reiter, B. (2025). AI in State Assessments. PowerPoint slides presentation, Hawai'i Department of Education. Presented at the National Assessment Governing Board quarterly board meeting.
- Tan, B., Armoush, N., Mazzullo, E., Bulut, O., and 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.