

Evidence-Centered Design for AI-Driven Automated Item Generation in Applied Mathematics

Edith Aurora Graf¹, Maria Elena Oliveri², Giulia Oliveri³, Emily Proctor²

¹Proof-of-Concept Learning LLC, ²Purdue University, ³Lincoln Southwest & Nebraska Wesleyan University

Abstract

We discuss how human-led Evidence-Centered Design (ECD) can inform the development of AI-supported automated item generation (AIG), as well as how it can be used as a quality-control mechanism. In this particular application, we describe how we collaborated with Claude Sonnet 5 to produce kinematics items together with interactive tools through an AIG pipeline. A preliminary qualitative analysis of a small number of generated items suggested they have identifiable strengths and weaknesses, which we discuss in depth per item.

1 Introduction

Automated item generation, or AIG (Irvine and Kyllonen, 2002; Gierl and Haladyna, 2012) has existed for many years, without the support of artificial intelligence (AI). Historically, AIG has been accomplished through the use of templates, which generate items with limited contextual and linguistic variation. As AI supports AIG and scoring at scale (A. A. von Davier et al., 2024), grounding these technologies in a cognitively-based theoretical framework is critical – here we consider AIG for open-ended mathematical modeling tasks, which are underrepresented in current AI applications.

Evidence-Centered Design (ECD; Mislevy et al., 2002; Mislevy et al., 2003) and Assessment Engineering (AE; Luecht, 2012; Luecht, 2025) can guide the creation of item models (Bejar, 2002), or schemas for AIG (Singley and Bennett, 2002). Huff et al. (2012) and Graf and Bejar (submitted) suggested that ECD task and evidence models can inform AIG, which in turn can be used to generate both tasks and the rubrics to score them. Graf and Flor (in press) used a hybrid, two stage AI-

supported AIG process to produce mathematics word problems (MWPs) from a schema. In the first stage, they established symbolic variables and constraints in accordance with the schema. In the second stage, they used an LLM to provide real-world contexts for the schema defined in the first stage, and to generate the MWPs. The intention of the first stage of the process was to reduce the incidence of inaccuracies that may be encountered when LLMs are solely responsible for generating mathematical content. The purpose of the second stage of the process was to generate items with sufficient contextual and linguistic variation, which is usually lacking among items generated from templates.

The use of human-led ECD and assessment engineering to guide design together with AI-supported AIG for item production has the potential to result in better assessment tasks than either humans or AI might generate alone. In the context of decision making, Hemmer et al. (2025) demonstrated that when there are asymmetries in either the information or capabilities contributed by a human and an AI collaborating as a team, the team works in a complementary manner, resulting in a better outcome than either human or AI would produce working alone. It seems likely that human and AI strengths would complement each other in additional domains, such as problem-solving, project work, or assessment development.

In addition to guiding task design, ECD can serve as a quality control mechanism for human-in-the-loop AI systems, with the potential to deliver enhanced validity, fairness, cost-efficiency, and pedagogical relevance compared to unanchored AI approaches in applied mathematics assessment.

Beichner (1994) developed and validated a multiple-choice test designed to assess students' understanding of kinematics graphs. The items

were designed in accordance with seven main objectives. Our goal in this research is to use ECD as a guiding framework to inform the development of constructed-response tasks and interactive tools. The tools are integrated with an AI-supported AIG pipeline, so that multiple tasks can be delivered, for practice or assessment purposes. We focus on two of the objectives identified by Beichner (1994 Table 1, p. 752):

- Given a velocity vs. time graph, the learner will determine acceleration
- Given a velocity vs. time graph, the learner will determine displacement

2 Approach

2.1 ECD Specification

Consistent with ECD, our first step was to consider which claims we wished to make as part of the assessment, the nature of evidence that would support those claims, and the tasks that would elicit that evidence. Since there was existing literature on the topic, the claims in this case were based on Beichner’s objectives. Since this is a preliminary exercise, the claims are deliberately narrow (we selected only two of Beichner’s seven objectives). The claims, evidence, tasks, and interactive tools are specified in Table 1.

2.2 Working with Claude to Produce the Interactive Tools

The request to build the interactive tools together with word problems began with this prompt to Claude Sonnet 5 (Anthropic, 2026):

“Let’s say I gave you some graphs, generated from R as before. In addition to wrapping word problem context around them, could you produce code that would render each graph in a web page, together with interactive tools that allowed students to find the slope of the tangent line at a point, or the area under a curve over an interval? Say, for example, I provided a velocity vs. time graphs and asked for the displacement between two time points, or a velocity vs. time graphs and asked to find the acceleration at a certain time point.”

Note that Claude had context for understanding this request, because the first author had engaged in earlier chats with Claude in which she provided

Table 1: Claims, Evidence, Tasks, and Interactive Tools for Velocity-Time Graphs

Claims	Evidence	Tasks	Interactive Tools
Given a velocity-time graph and supporting interactive tools, the learner can find instantaneous acceleration at a particular time point (Beichner’s Objective #2)	The learner is able to identify the time point in question, and recognize that the instantaneous acceleration corresponds to the slope of the tangent line to the velocity-time graph at that point.	The learner is given a contextualized velocity-time graph, a time point, and asked to find the instantaneous acceleration at that point. The learner has access to two interactive tools.	As the learner drags or sets a time point on the graph, the tangent line and its slope at that time point are displayed.
Given a velocity-time graph and supporting interactive tools, the learner can find displacement between two particular time points (Beichner’s Objective #3)	The learner is able to identify the two time points in question, and recognize that the displacement corresponds to the net area between the velocity-time graph and the horizontal axis between those two time points.	The learner is given a contextualized velocity-time graph, two time points, and asked to find the displacement between those two time points. The learner has access to two interactive tools.	As the learner drags or sets the time points on the graphs, the area(s) between the curve and the horizontal axis together with their values for that interval are displayed.

kinematics graphs generated from R (R Core Team, 2026) and asked Claude to wrap word problem context around them.

Claude delivered a test html file for inspection that included both a couple of items and the interactive tools. At this point, several rounds of iteration between the first author and Claude ensued to modify the tools. These were primarily UI/UX modifications. For example, although the tangent line in the tangent tool is a different color than the graph, it originally completely obscured the graph when over a constant or linear segment. It was requested that the tangent line be made dashed or slightly transparent, so that both the graph and the tangent line would always be visible.

2.3 Working with Claude to Integrate the Interactive Tools with an AIG Pipeline

The next goal involved combining the interactive tools with multiple items generated via AIG. In contrast to earlier work, Claude produced all the

code needed for the AIG pipeline. This request was initiated with the following prompt:

“Thanks. Ideally, I'd like to use this as part of an AIG workflow. In other words, rather than writing separate code for each item that then gets rendered in a web page, I'd like to request some number of items (say, 20), have the parameters that create the graphs randomly determined (but without repeat items), and also have the word problem context added in). Then I port the whole thing to the multi-page website. Before you start though, let's explore the feasibility of this and how it would work.”

In the chat that followed, the first author and Claude discussed design decisions for the pipeline, and how the constraints on randomization should be defined. Claude produced all the necessary R scripts, which were gathered in a single directory. Again several rounds of iteration took place.

To initiate the pipeline, a user sources an item features R script, and specifies both the number of items to generate and a unique seed. Three files are produced that describe the item features, including the mathematical parameters for the graphs. The csv file, which contains a subset of the features, is for human review. The JSON file is uploaded to the AI chat, together with processing instructions. The rds file is used downstream in the pipeline.

The next step is to provide the JSON file in the AI chat together with processing instructions. The AI takes the item feature information provided in the JSON file and produces scenario text encapsulated in an R file. Note that the scenarios are intended to be plausible given the item features, but no item feature information is contained in the R file that contains the scenario text.

The user then sources an R script which merges item feature information (this time from the rds file) and the scenario text into another JSON file, which includes all item information.

Finally, the JSON file containing all item information is laid out according to an html template, and all the generated items together with the interactive tools are captured as a single html file. Note that while the item features are recoverable using the seed, the scenario text is not reproducible and so either the JSON file that includes all item information or the final html file must be saved.

The AIG pipeline is summarized in Figure 1. Note that there is a clear division of labor: The

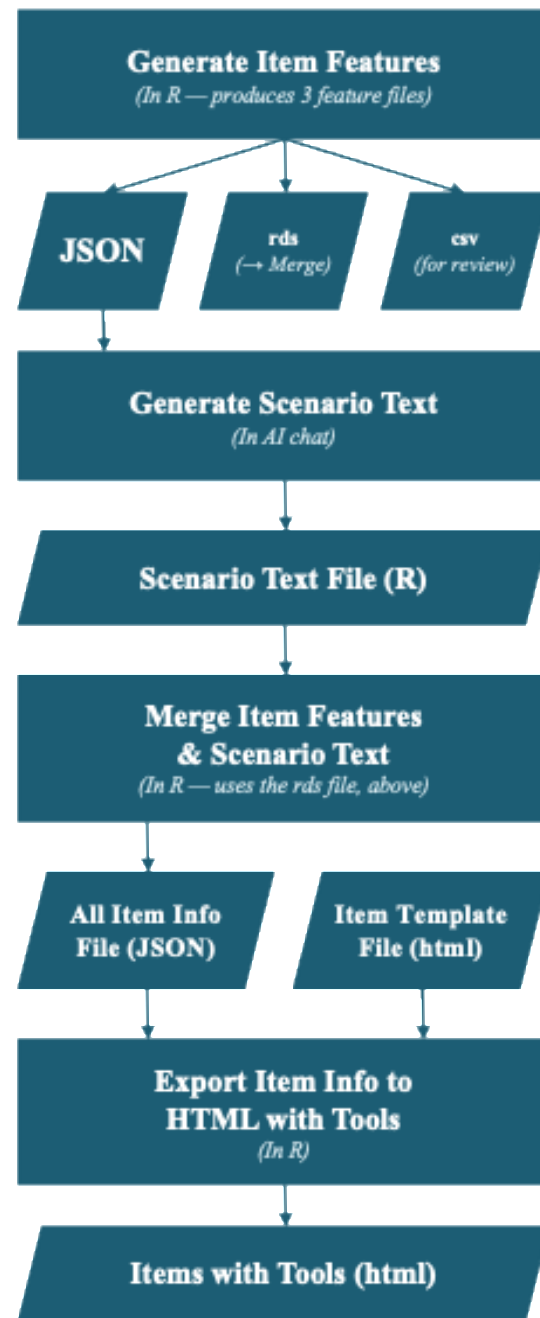


Figure 1: AIG Pipeline with Interactive Tools

mathematical properties of the items are generated by R, while the LLM is responsible for generating the scenario text to accompany the graphs. This is an example of application of the “divide and conquer” technique (von Davier, 2024; von Davier, 2025).

2.4 A Run through the AIG Pipeline

Twenty graphs were generated with two questions each (one requesting acceleration and the other displacement). An example is shown in Figure 2.

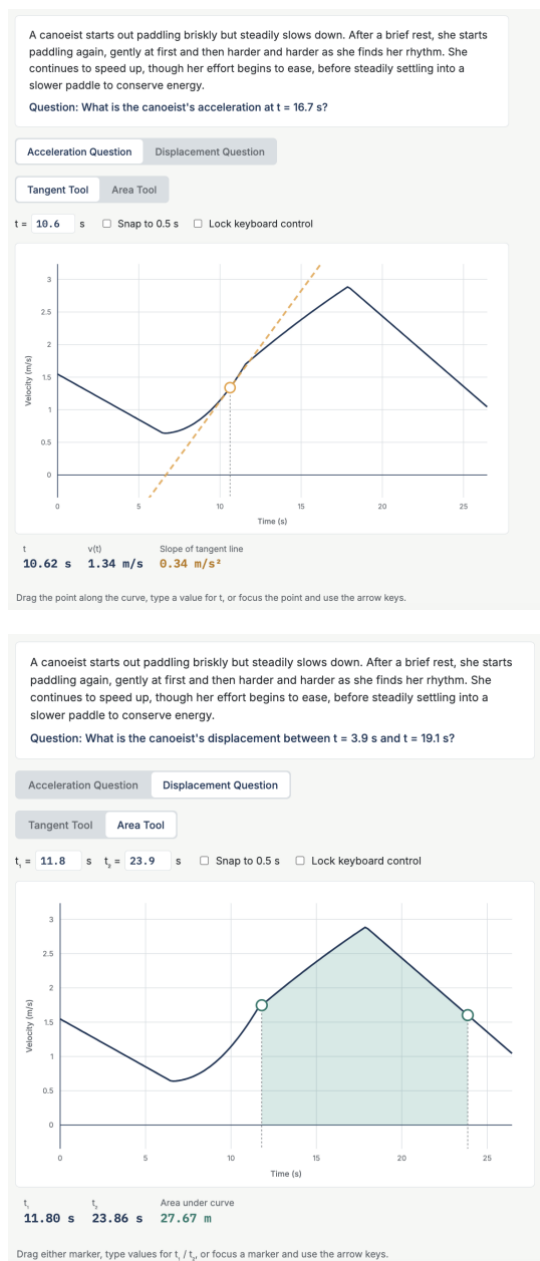


Figure 2: The Paddling the Lake Item with Two Questions and the Interactive Tools

As mentioned, several rounds of iteration took place following development of the pipeline scripts. For example, the layout was initially designed so that only the slope of the tangent tool was available for the acceleration questions, and only the area under the curve tool was available for the displacement questions. However, learners sometimes perform slope calculations when area calculations are needed (Beichner, 1994), and the reverse is also possible. It's helpful at this point in our discussion to review Table 1 again, which states that we are interested in evidence that the learner can "...recognize that the instantaneous

acceleration corresponds to the slope of the tangent line..." and the learner can "...recognize that the displacement corresponds to the net area..." However, as the layout was initially designed, the learner had no opportunity to demonstrate evidence to the contrary, illustrating the misconception identified by Beichner.

This is an example where ECD is potentially very valuable as a quality control mechanism: Once items are generated, it is useful to ask the question: Do these items really elicit the target evidence, or do they not allow that evidence to be collected? Asking that question led to continued conversations with Claude, during which it was discussed whether it would be better to decouple tool availability from question type. In the end, it was resolved to provide the learner with a choice between two modes when initially loading the page: practice mode or assessment mode. In practice mode, tool availability is determined by question type: Only the tangent tool is available when acceleration is requested, and only the area tool is available when displacement is requested. Also, scaffolding is present via more explicit labeling. In assessment mode (which is shown in Figure 2), both tools are available for both question types (the correct choices have been made in Figure 2).

3 Analysis

In this section, we present qualitative analysis of four of the 20 generated items. In earlier work (Graf et al., submitted) we developed a draft instrument to evaluate the quality of kinematics tasks produced using AI-supported AIG. The instrument spans five binary dimensions, including Accuracy, Completeness, Plausibility, Familiarity, and Language. Accuracy indicates whether the scenario text is a mathematically accurate description of the motion described by the graph. Completeness refers to whether sufficient information is supplied by the question to find the answer. Plausibility indicates whether the scenario text is realistic, and Familiarity whether it is sufficiently familiar to the intended audience. Finally, Language refers to whether the vocabulary and complexity of sentence structure is suitable for the intended audience.

In our earlier work, 20 generated kinematics tasks were independently scored by two expert raters and an LLM using the instrument. In addition to providing a score for each item on each

dimension, raters provided a justification. Agreement was uniformly low between the two raters and between each of the raters and the LLM. In the case of the raters, this may have occurred because the instrument was newly-developed and the raters lacked sufficient training. But it also likely occurred because each rater sometimes noticed issues that the other had missed; but following discussion there was often agreement that it was indeed an issue. Another interesting pattern was that although agreement between the human raters was poor, bias was not consistently present. The LLM, however, tended to score more favorably than either of the human raters.

The written justifications for the scores proved more useful than the scores themselves, and raters reflected on them in arriving at a single set of reconciled scores. In this work we took a different approach: Instead of evaluating the quality of the generated items using a scoring instrument, two reviewers analyzed a total of four items qualitatively, making notes with respect to each dimension included in the instrument.

3.1 Paddling the Lake

The Paddling the Lake item is shown in Figure 2, and was evaluated by one reviewer. With respect to accuracy, the reviewer noted that the scenario text is ‘mostly coherent qualitatively,’ but that “‘effort” is not a kinematic quantity, so the mapping to acceleration remains interpretive.’ The item’s plausibility was rated highly, and the suitability of the language was deemed good, apart from some anthropomorphic description. The item was judged as incomplete, but it’s possible that only the scenario text (and not the graph and the prompt) were considered in making this judgment. The item was also considered familiar.

3.2 Missed the Exit

The Missed the Exit item is shown in Figure 3. With respect to accuracy, the sole reviewer noted some imprecision, however the most obvious issue concerns plausibility: For a motorist to back up on the highway due to missing an exit is both unsafe and illegal, as well as unlikely to occur. The suitability of the language was considered mixed due to the introduction of extraneous practical concerns that distract from 1-D motion. The item was judged as incomplete, but it’s possible that only the scenario text (and not the graph and the prompt) were considered in making this judgment.



Figure 3: The Missed the Exit Item with Two Questions and the Interactive Tools

Familiarity was deemed moderate overall, although it was commented that the on-ramp setting could be confusing and unfamiliar.

3.3 Open Highway

The sole reviewer of the Open Highway item (Figure 4) indicated that the item is mathematically accurate in that the description of motion in the scenario text correctly captures what is represented by the graph. Overall, the scenario was also considered plausible behavior for someone merging into traffic and responding to traffic conditions. Finally, the item was judged acceptable

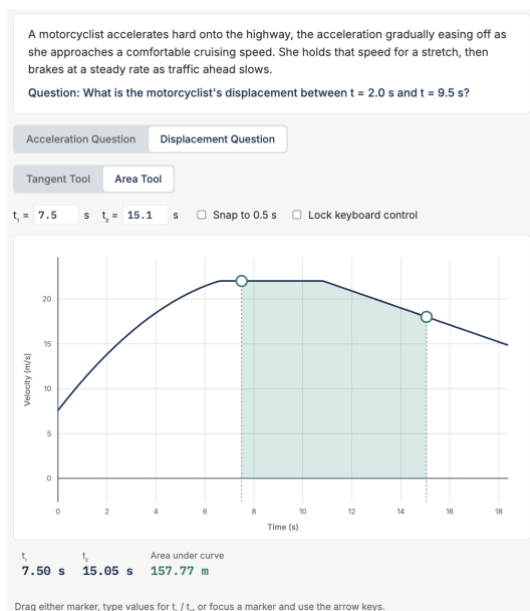


Figure 4: The Open Highway Item with Two Questions and the Interactive Tools

with respect to Completeness, Language, and Familiarity, as it discussed all aspects of motion without unnecessary jargon.

3.4 Backing Down the Driveway

The Backing Down the Driveway item, which was reviewed by two experts, is shown in Figure 5. With respect to accuracy, the first reviewer felt it is “coherent and clean,” while the second reviewer had concerns with the phrase, “Partway down, she shifts into drive and accelerates forward at a fairly

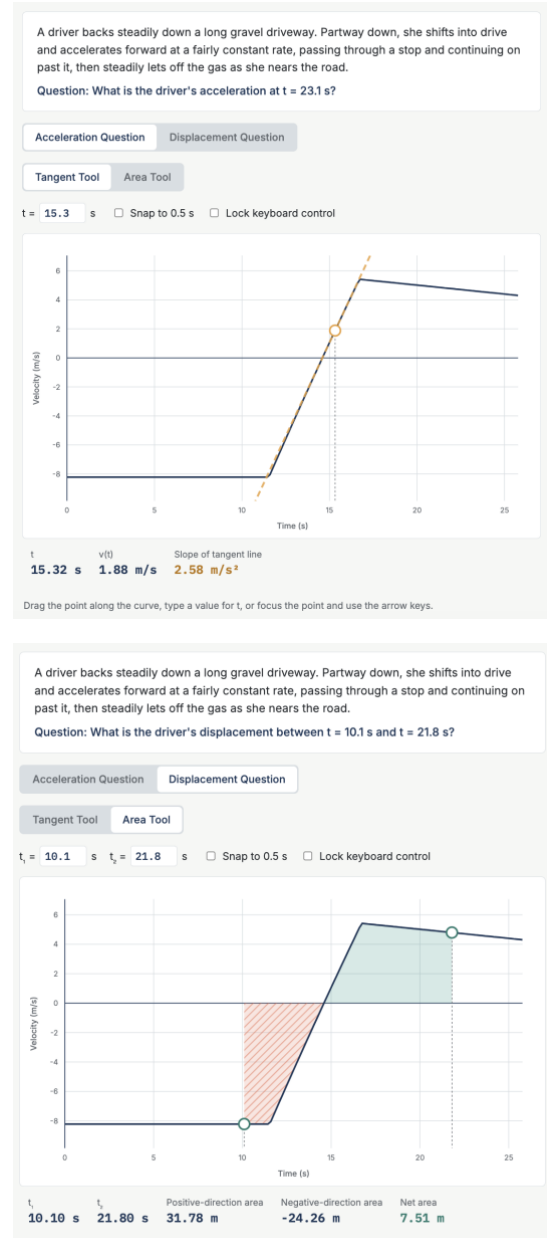


Figure 5: The Backing Down the Driveway Item with Two Questions and the Interactive Tools

constant rate.” While acknowledging that this is a technically accurate description since the second segment has positive slope, the reviewer pointed out that most humans don’t describe vehicles going backwards as accelerating forward. This certainly could be a point of confusion on an assessment that would likely introduce construct-irrelevant difficulty. In addition, the second reviewer thought that the phrase “passing through a stop and continuing on past it,” was unclear, because it could indicate ignoring a stop sign rather than transitioning from negative to positive velocity.

The first reviewer indicated that plausibility on this item is high. The second reviewer disagreed, arguing that the sentence, “Partway down, she shifts into drive and accelerates forward at a fairly constant rate,” is impossible, because to get the segment shown in the graph, the driver would need to slow down using the brake while in reverse before shifting into drive, all while maintaining constant acceleration. Reviewer 2 had additional plausibility concerns regarding the meaning of “stop.”

Neither reviewer felt the item is complete, though for different reasons. Reviewer 1 may have made this judgment based on the scenario text alone, whereas Reviewer 2 indicated that not all segments of the graph seemed to be represented in the scenario text. Regarding Language and Familiarity both reviewers were mostly positive, with Reviewer 2 again pointing out potential ambiguity about the meaning of the word “stop.”

3.5 Analysis Summary and Discussion

Although very preliminary, some interesting observations can be drawn from the present analysis. First, we saw examples of items that were more (Paddling the Lake, Open Highway) or less (Missed the Exit, Backing Down the Driveway) successful from the perspectives of two expert reviewers. The analysis is informative with respect to the nature of both the strengths and weaknesses of items produced using AI-supported AIG. With respect to strengths, the contexts do seem varied, often familiar, and potentially engaging. With respect to weaknesses, the tasks were sometimes unrealistic, or even described an unsafe and illegal scenario (Missed the Exit). Language that could be difficult for a human to interpret, potentially resulting in construct-irrelevant difficulty, was also observed.

Given that only four non-randomly selected items were studied, there’s a lot we can’t conclude. For example, we don’t know how likely the kinds of weaknesses we observed are to occur given a similar or even identical set-up. We can’t make claims about differences in accuracy and plausibility between items that involve a change in direction vs. those that do not, though this would be interesting to explore. A more complete analysis would involve reviewing a much larger pool of randomly generated items. As noted, we have done something like this in the past, using an instrument

to score according to five attributes of item quality (Accuracy, Plausibility, Completeness, Language, and Familiarity). In future work we might refine this instrument, or we might approach evaluation as we did here: Reviewers take notes using these attributes as a guide. Subsequently we could analyze those notes, perhaps using thematic analysis (Boyatzis, 1998).

A practical question involves how to mitigate the item weaknesses. One solution is to rewrite or discard items with weaknesses. Of course human review is always essential. But providing more guidance might help reduce the number of items from the AIG pipeline that need to be rewritten or discarded. The first author had a discussion with Claude about this, in which Claude suggested providing a review checklist relevant to the domain of interest. For example if a scenario involves motion of a train, it is plausible the train will reverse direction by going backwards after reaching the end of the line. If the scenario involves motion of a runner or a cyclist, however, it is far more likely the runner or cyclist will turn around to reverse direction. An LLM could be prompted to review for consistency between the way the motion is described and the type of vehicle involved. Similarly, it could be prompted to disambiguate meaning when using domain-specific terms that might be interpreted colloquially.

Another approach that might be effective for mitigating the weaknesses would be to reverse the steps in the pipeline: Rather than generating graphs that the LLM must provide scenarios for, the LLM could be prompted to provide scenarios together with some corresponding numeric ranges. R could then sample from within those ranges to produce the graphs, which would be combined with the scenario text as before. The approach is similar in that the LLM is still generating the scenario text and R is still producing the graphs and the tools, but the scenarios inform the graphs rather than the other way around.

4 Discussion and Conclusion

In this paper we described how human-led ECD can be used to inform the development of an AI-supported, AIG pipeline for kinematics tasks that incorporate interactive tools. Claude was used at every step of the pipeline development, and programmed all of the R scripts. However, development was truly collaborative. Multiple iterations between a human and Claude were used

to refine the tools, the questions, and the user interface.

We discussed an example of how in addition to guiding AI-supported AIG, ECD can act as a quality control mechanism. When it was recognized that question type (acceleration vs. displacement) and tool type (tangent tool vs. area tool) were confounded, and that this was not ideal for surfacing a potential misconception, the interface was redesigned to allow for both practice and assessment modes.

Certainly a strength of collaborating with AI on a project like this is the speed with which interactive tools can be developed and prototyped, a process which would ordinarily take weeks or months, depending on the complexity of the project. Note here however that only a front end was built: The interactive tools are not designed to collect any data.

The AIG pipeline used R to produce the item features (including the mathematical parameters for the graphs), and the LLM to produce the scenario texts. The item features and scenario texts were then merged at a later stage to finalize the items, which were rendered in a web page. The objective was to leverage the complementary strengths of each technology.

We then conducted a preliminary qualitative analysis consisting of an in-depth examination of a very small number of generated items with respect to five quality attributes (Accuracy, Plausibility, Completeness, Language, and Familiarity). Based on this analysis we cannot make any claims about how frequent certain item strengths or weaknesses are likely to be. We can, however, use the analysis to assert the existence of strengths and weaknesses, which we discussed in detail for each item analyzed. Next steps for the analysis work involve evaluating larger numbers of randomly-generated items, either using a refined version of the instrument developed as part of our earlier work or using reviewer notes followed by thematic analysis.

An obvious next question is how we can mitigate the item weaknesses, so fewer generated items need to be rewritten or discarded. One possibility was suggested by Claude: Create a checklist that defines real-world contingencies that are important to consider. The other is to reverse the order of the AIG pipeline: Instead of providing item features that an LLM must use as constraints in defining a scenario, have the LLM generate the

scenarios first, together with numerical ranges from which R can randomly sample to produce graphs. Which pipeline order might yield a greater proportion of retained items? It suggests an experiment to find out.

References

- Anthropic. 2026. Claude Sonnet 5. Large language model.
- R. J. Beichner. 1994. Testing student interpretation of kinematics graphs. *American Journal of Physics*, 62(8):750–762.
- Isaac I. Bejar. 2002. Generative testing: From conception to implementation. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*, pages 199–217. Lawrence Erlbaum Associates, Mahwah, N.J.
- Richard E. Boyatzis. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE. Google-Books-ID: _rfCIWRhIKAC.
- Mark J. Gierl and Thomas M. Haladyna, editors. 2012. *Automatic Item Generation: Theory and Practice*. Routledge. Google-Books-ID: llaU71hlvFEC.
- Edith A. Graf and Isaac I. Bejar. submitted. The impact of evidence-centered design on item modeling and automated item generation: Past, present, and future. In Eric M. Tucker and Maria Elena Oliveri, editors, *Modeling What Matters: The Research and Legacy of Robert J. Mislevy*.
- Edith A. Graf and Michael Flor. in press. Automated generation of mathematics word problems by a symbolically-constrained LLM. In von Davier, Alina A. and Yan, Duanli, editors, *Artificial Intelligence in Educational Learning and Assessment*.
- Edith A. Graf, Cheryl L. Lizano, and Maria Elena Oliveri. submitted. Principled Automated Content Generation for Mathematical Problem-Posing Tasks.
- Patrick Hemmer, Max Schemmer, Niklas Köhl, Michael Vössing, and Gerhard Satzger. 2025. Complementarity in human-AI collaboration: concept, sources, and evidence. *European Journal of Information Systems*, 34(6):979–1002.
- Kristen Huff, Cecilia B. Alves, James Pellegrino, and Pamela Kaliski. 2012. Using Evidence-Centered Design Task Models in Automatic Item Generation. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, pages 102–118. Routledge.

Sidney H. Irvine and Patrick C. Kyllonen, editors. 2002. *Item generation for test development*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Richard M. Luecht. 2012. An introduction to assessment engineering for automatic item generation. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation*. Routledge.

Richard M. Luecht. 2025. *Assessment Engineering in Test Design: Methods and Applications*. Routledge, New York.

Robert J. Mislevy, Linda S. Steinberg, and Russell Almond. 2002. On the roles of task model variables in assessment design. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*. Lawrence Erlbaum Associates, 0 ed.

Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond. 2003. Focus article: on the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1):3–62.

R Core Team. 2026. R: A language and environment for statistical computing.

Mark K. Singley and Randy E. Bennett. 2002. Item generation and beyond: applications of schema theory to mathematics assessment. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*. Lawrence Erlbaum Associates.

Alina A. von Davier. 2024. Pre-conference workshop at International Association for Computerized Adaptive Testing (IACAT). In Seoul, South Korea.

Alina A. von Davier. 2025. Pre-conference workshop at National Council on Measurement in Education (NCME). In Denver, CO, United States.

Alina A. von Davier, Andrew Runge, Yena Park, Yigal Attali, Jacqueline Church, and Geoff LaFlair. 2024. The item factory: intelligent automation in support of test development at scale. In Hong Jiao and Robert W. Lissitz, editors, *Machine Learning, Natural Language Processing, and Psychometrics*, page 0. Emerald Publishing Limited.