Pre-Pilot Optimization of Conversation-Based Assessment Items Using Synthetic Response Data

Tyler Burleigh
Khan Academy
tylerb@khanacademy.org

Jing Chen
Khan Academy
jing@khanacademy.org

Kristen DiCerbo
Khan Academy
kristen@khanacademy.org

Abstract

Correct answers to math problems don't reveal if students understand concepts or just memorized procedures. Conversation-Based Assessment (CBA) addresses this through AI dialogue, but reliable scoring requires costly pilots and specialized expertise. Our Criteria Development Platform (CDP) enables pre-pilot optimization using synthetic data, reducing development from months to days. Testing 17 math items through 68 iterations, all achieved our reliability threshold (MCC $\geq$ 0.80) after refinement – up from 59% initially. Without refinement, 7 items would have remained below this threshold. By making reliability validation accessible, CDP empowers educators to develop assessments meeting automated scoring standards.

1 Background

When students solve math problems correctly, teachers face a critical challenge: they cannot tell if students understand the concepts or just memorized the steps. A student who correctly solves $1.5(2-4h) = 6h$ might understand why division maintains equality, or might simply execute a memorized procedure. When students do not solve a problem correctly, the only information available is that they entered an incorrect answer. It is unknowable whether they had a partial or incomplete understanding of the problem. Traditional tests cannot provide evidence about students' thought process when answering questions, creating a gap that affects teaching decisions and student support.

Conversation-Based Assessment (CBA) enables the assessment of conceptual understanding through adaptive dialogue (Yildirim-Erbasli and Bulut, 2023). In CBA, students explain their reasoning, similar to constructed response items (Williamson et al., 2012). Unlike static written responses, CBA adapts based on student answers – asking follow-ups when needed and providing appropriate feedback (Jackson et al., 2018). This interaction provides evidence indicating whether students grasp underlying concepts.

CBA technology has progressed from scripted to generative systems. Early approaches required authoring specific response-reply pairs (Zapata-Rivera et al., 2015), essentially building complete dialogue trees that anticipated every possible student response. Later systems like Quizbot used semantic matching to map student responses to pre-written feedback (Ruan et al., 2019), but educators still had to design and construct all potential conversation paths beforehand.

Large Language Models (LLMs) introduced in 2022 (OpenAI, 2024) marked a paradigm shift. Instead of pre-building dialogue trees, these newer systems allow students to respond openly in their own words, with the AI using NLP methods to understand and categorize responses dynamically (Bergerhoff et al., 2024). This eliminates the burden of anticipating and scripting every possible conversation branch, making CBA development accessible to educators without the resources for complex dialogue engineering.

Yet this freedom from pre-coding dialogue paths creates a different challenge. When systems can accept any student response rather than matching against predetermined patterns, they must interpret novel expressions of understanding in real-time. While these models are capable of such evaluation, without explicit guidance about what constitutes conceptual mastery, their scoring decisions may lack the consistency needed for reliable assessment.

Scoring criteria provide this needed guidance, giving structure to open-ended evaluation. By explicitly defining what constitutes conceptual mastery for each item, these criteria enable CBA systems to evaluate diverse student responses consistently and generate appropriate follow-ups. Good criteria help AI Scorers match human grader reliability (Henkel et al., 2024), especially when subject

October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

matter experts write item-specific criteria rather than generic prompts (Frohn et al., 2025).

Creating reliable scoring criteria requires meeting established assessment standards, with AI and human graders reaching similar conclusions. Educational assessment typically requires strong agreement (e.g., $\kappa \geq 0.70$) (Williamson et al., 2012; Wood et al., 2021). These thresholds challenge traditional empirical validation because they require extensive time and resources to reach (Williamson et al., 2012). Developers draft criteria, pilot them with real students, and compare AI scores to human ratings. Discrepancies trigger revision and re-piloting. Most items require multiple cycles, taking months and requiring fresh student data each time.

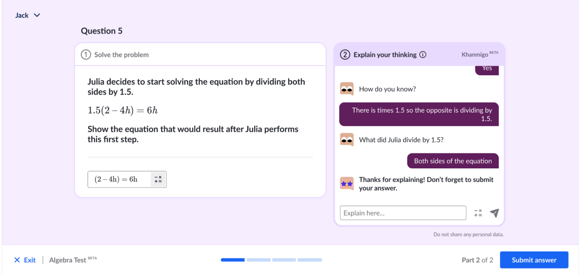
Even when time and resources are available, the validation process requires specialized technical knowledge that content authors often lack, such as: dataset development (developing and labeling balanced, diverse synthetic datasets), metric computation (choosing and calculating coefficients), iteration management (managing multiple criteria refinements and their associated datasets), and interpreting results (setting targets and identifying which changes were meaningful). Without this expertise, efforts may yield unreliable results.

These twin challenges – lengthy validation cycles and specialized expertise requirements – create a bottleneck in CBA development. Without tools to test criteria before student pilots, developers must choose between deploying potentially unreliable assessments or investing months in iterative pilot studies. At Khan Academy, these challenges drove us to develop an alternative to time-consuming student pilots for validating scoring criteria. To solve this problem, we developed a platform that lets creators test criteria using synthetic data and provides step-by-step guidance.

1.1 Explain Your Thinking (EYT): A Modern Conversation-Based Assessment System

Before describing our solution, we first describe the EYT system itself. Understanding how EYT uses criteria to both score responses and generate follow-up questions reinforces why criteria quality is so critical to CBA success.

Explain Your Thinking (EYT) is our implementation of modern CBA. Students first solve problems, then explain their reasoning in AI-guided conversations. For example, when a student solves $1.5(2-4h) = 6h$ by dividing both sides by 1.5, we know whether or not they can execute the procedure. But EYT goes deeper: can they explain why division maintains equality? Do they understand that division and multiplication are inverse operations? The system uses scoring criteria to evaluate these conceptual understandings and generate appropriate follow-up questions.



Julia decides to start solving the equation by dividing both sides by 1.5.

$1.5(2-4h) = 6h$

Show the equation that would result after Julia performs this first step.

$(2-4h) = 6h$

Yes

How do you know?

There is times 1.5 so the opposite is dividing by 1.5.

What did Julia divide by 1.5?

Both sides of the equation

Thanks for explaining! Don't forget to submit your answer.

Figure 1: Screenshot of the Explain Your Thinking conversation-based assessment item type. The student first answers a math problem (left), and then has a conversation about the problem (right) which is designed to assess their conceptual understanding.

Each assessment activity starts with a math problem, the student's answer, and criteria defining complete understanding. The platform operates through three integrated functions that enable assessment. First, it recognizes varied expressions of concepts, allowing students to explain their thinking in their own words. Second, it generates probing questions that explore understanding without revealing answers. Third, it maintains assessment validity by avoiding teaching during the evaluation process.

Importantly, EYT's effectiveness depends on a criteria-driven cascade. At each turn, an AI Scorer evaluates the conversation history to determine which criteria the student has satisfied. These evaluations then flow to a Response Generator, which receives a list of unsatisfied criteria and generates targeted follow-up questions to probe those specific gaps. When criteria are vague or missing, this cascade breaks down: the AI Scorer misclassifies responses, passing incorrect information downstream, and the Response Generator asks about the wrong concepts, leading to unproductive conversations.

Students experience a natural conversation flow. They explain their approach and receive targeted follow-ups that probe gaps without teaching. The conversation continues until students demonstrate understanding or reach a four-turn limit.

2 Criteria Development Platform (CDP)

Given that poor criteria can derail EYT's assessment cascade and compromise validity, we needed a way to ensure criteria quality before deployment. Our Criteria Development Platform (CDP) addresses this need by enabling content creators to test and refine AI scoring criteria using synthetic student responses, eliminating the months-long pilot cycles traditionally required for validation.

CDP operates through an iterative workflow where creators write scoring criteria, generate synthetic responses that test edge cases, evaluate AI Scorer performance against these responses, and refine their criteria based on the results.

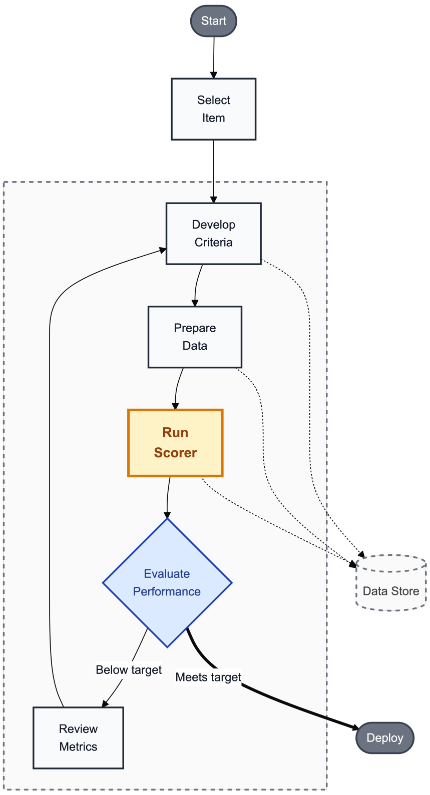


Figure 2: The Criteria Development Platform's iterative workflow. Content creators write scoring criteria, generate synthetic responses, test performance, and refine their criteria based on results.

CDP addresses both core challenges of criteria development. First, it reduces validation time from months to days by eliminating the need for multiple rounds of student pilots. Creators can test dozens of iterations in hours or days rather than weeks or months. Second, it provides guided support that makes reliable assessment creation accessible without specialized expertise. The platform automatically tracks versions, computes performance metrics, and provides targets and actionable feedback to guide criteria development.

To evaluate CDP's effectiveness, we analyzed 68 development cycles from six content creators developing 17 math assessment items. Our analysis addresses three key research questions:

1. **Engagement**: Do content creators effectively engage in iterative refinement when using CDP?
2. **Improvement**: When creators iterate, do their scoring criteria demonstrate measurable performance gains?
3. **Achievement**: What proportion of items ultimately meet established reliability standards?

The following sections detail CDP's design and demonstrate its effectiveness through empirical analysis of these development cycles.

2.1 How CDP works

Creators follow four steps (Figure 2) to create scoring criteria. They select an item and then iterate through: writing criteria, generating data, and testing performance until meeting standards. Throughout this cycle, the tool preserves all data and metrics, allowing creators to track improvements and learn from each iteration. This four-step process addresses the challenges of time and expertise: synthetic data allows rapid iteration, while metric generation and feedback provide scaffolding for creators.

2.1.1 Step 1: Selecting the item

Content creators start by selecting an item for which to develop criteria.

2.1.2 Step 2: Writing scoring criteria

Next, creators write up to seven criteria that define a complete response. For instance, an item about solving equations might include criteria like "identifies the inverse operation needed" and "explains why division undoes multiplication." The platform tracks all versions of criteria, allowing creators to try different approaches and revert to previous versions as needed.

2.1.3 Step 3: Generating test data

Creators need synthetic data to evaluate their criteria without real students. The platform guides creators in developing balanced datasets of 150 simulated responses per test, with 50 responses each from correct, partially correct, and incorrect categories. Each response includes the student's initial answer, their conversational explanation, and a human-assigned ground truth label (correct, partially correct, or incorrect) indicating the response's category. Content creators must carefully assign these ground truth labels when developing the synthetic dataset, as they serve as the authoritative reference for evaluating the AI Scorer's performance. This balanced distribution across the three categories ensures comprehensive testing of the criteria. The sample size of 150 was determined through simulation-based power analysis, achieving >80% Bayesian posterior probability that MCC ≥ 0.8 when the true MCC is at least 0.84. This provides strong statistical evidence for identifying scorers that meet the performance threshold.

Creators generate these responses through a combination of manual writing and AI assistance. To ensure quality and authenticity, we instructed creators to manually write at least 10-15 example responses for each correctness category, capturing realistic student thinking patterns. (Note that for scoring purposes, the AI Scorer uses a binary classification approach and treats partially correct responses as incorrect. However, including partially correct responses in the dataset serves a critical purpose: they enhance diversity by capturing edge cases and boundary conditions where students demonstrate some but not all required understanding. This helps creators test whether their criteria can distinguish between complete and incomplete responses, identifying potential ambiguities before deployment.)

When using the AI generator, the platform prompts a language model such as GPT-4.1 with these manually-created examples, information about the item, and the criteria. The model generates unique, plausible student responses matching the specified category. It receives instructions to vary both reasoning patterns and writing style. This ensures responses remain meaningfully different from the provided examples. Creators must verify all AI-generated responses and correct ground truth labels if necessary before adding them to their dataset.

This approach combines human expertise with AI's ability to generate variations at scale. Human creators identify realistic student thinking patterns. AI generates diverse examples based on these patterns. Human oversight ensures classification accuracy throughout. We also encouraged creators to include edge cases in test data, such as correct reasoning with unusual terminology. Testing against these challenging cases helps creators identify and fix ambiguities in their criteria before real students encounter them.

The system helps ensure response diversity through similarity checking. When generating synthetic responses, the platform uses semantic embeddings to compare each new response against all other responses within the same correctness category, flagging pairs that exceed 85% similarity. For mathematical problems, the similarity checker includes additional detection for responses that differ only in numerical values, as these may have high semantic similarity despite representing fundamentally different solutions. This prevents dataset contamination from near-duplicate responses that would artificially inflate performance metrics.

2.1.4 Step 4: Testing and refining

With criteria and test data ready, creators run the AI Scorer and see how well it performs with the current criteria. The platform calculates performance metrics by comparing the AI Scorer's predictions against the human-assigned ground truth labels from the synthetic dataset. Metrics like accuracy and false positive rates reveal how well the AI Scorer aligns with human judgment, helping creators identify problems with their criteria definitions. Additionally, the tool provides the AI Scorer's reasoning for each evaluation, showing where criteria might be unclear or ambiguous (Figure 3). This transparency addresses the expertise challenge by making the AI Scorer's evaluation process interpretable to non-experts.

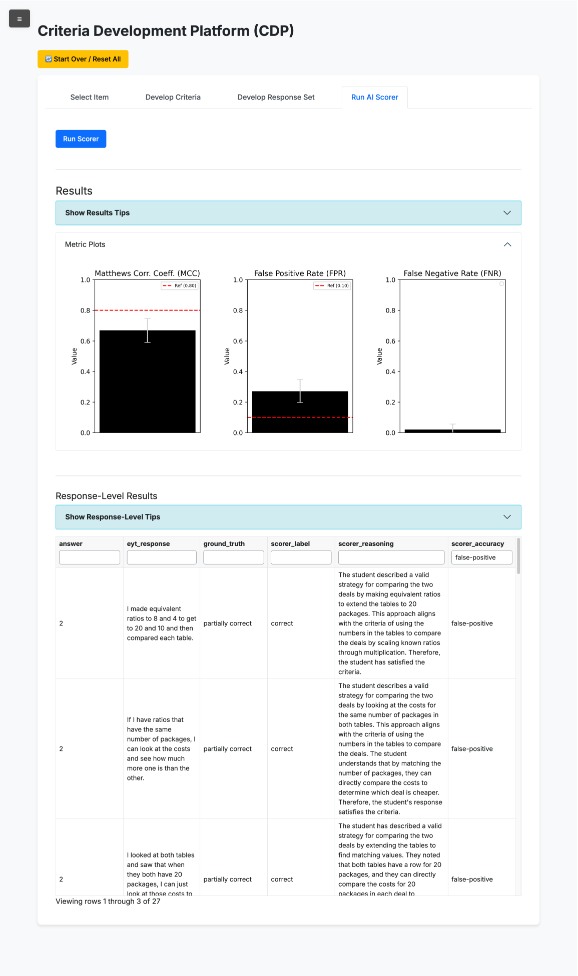
3 Research

To evaluate CDP's effectiveness in enabling prepilot optimization, we conducted an empirical analysis of platform usage data. This analysis directly addresses the three research questions posed in our Aims: examining creator engagement patterns, measuring performance improvements, and determining achievement rates.

3.1 Methods

3.1.1 Dataset

Six content creators (curriculum specialists and assessment designers) independently developed 17 mathematics items over three months. Each development cycle followed the same workflow: writing scoring criteria, generating synthetic responses, and evaluating AI Scorer performance. Items covered grades 6-12 mathematics, addressing algebra,



Criteria Development Platform (CDP)

Results

Show Results Tips
Metric Plots
Matthews Corr. Coeff. (MCC)
Ref (0.80)
False Positive Rate (FPR)
Ref (0.10)
False Negative Rate (FNR)
Value
1.0
0.8
0.6
0.4
0.2
0.0

Response-Level Results

answer	eyt_response	ground_truth	scorer_label	scorer_reasoning	scorer_accuracy
					false-positive
2	I made equivalent ratios to 8 and 4 to get to 20 and 10 and then compared each table.	partially correct	correct	The student described a valid strategy for comparing the two deals by making equivalent ratios to extend the tables to 20 packages. This approach aligns with the criteria of using the numbers in the tables to compare the deals by scaling known ratios through multiplication. Therefore, the student has satisfied the criteria.	false-positive
2	If I have ratios that have the same number of packages, I can look at the costs and see how much more one is than the other.	partially correct	correct	The student describes a valid strategy for comparing the two deals by looking at the costs for the same number of packages in both tables. This approach aligns with the criteria of using the numbers in the tables to compare the deals. The student understands that by matching the number of packages, they can directly compare the costs to determine which deal is cheaper. Therefore, the student's response satisfies the criteria.	false-positive
2	I looked at both tables and saw that when they both have 20 packages, I can just look at those costs to	partially correct	correct	The student has described a valid strategy for comparing the two deals by extending the tables to find matching values. They noted that both tables have a row for 20 packages, and they can directly compare the costs for 20 packages in each deal to	false-positive

Viewing rows 1 through 3 of 27

Figure 3: The Criteria Development Platform's AI Scorer interface displaying performance metrics (MCC, FPR, FNR) and response-level results with the AI's scoring reasoning for each evaluation.

geometry, and ratio topics aligned with Common Core State Standards.They generated 68 development cycles. When content creators tested the same criteria version multiple times during development, we included only the final run for each version in our analysis. This resulted in 61 distinct criteria versions with 10,200 synthetic response evaluations.

We analyzed two distinct item groups. Eight items (47%) underwent iterative refinement through multiple criteria revisions. Nine items (53%) achieved strong performance without criteria changes, maintaining consistent criteria across runs. This division lets us examine both the refinement process and cases of immediate success.

3.1.2 Performance Metrics

We evaluated AI Scorer performance using four metrics that measure agreement between AI-generated scores and ground truth labels (expert human scoring):

Matthew's Correlation Coefficient (MCC) serves as our primary metric for evaluating scoring reliability, considering all classification outcomes. Values range from -1 to +1, with ≥ 0.80 threshold. MCC balances imbalanced datasets.

Three additional metrics provide comprehensive evaluation alongside our primary MCC metric:

Cohen's kappa (κ_C) quantifies agreement beyond what random chance would produce, with ≥ 0.81 indicating substantial reliability (per Landis and Koch (1977)).

False Positive Rate (FPR) tracks a critical failure mode: marking incorrect responses as correct, which terminate conversations prematurely. Our threshold of ≤ 0.10 ensures the AI does not often terminate conversations prematurely.

Accuracy: proportion correct.

3.1.3 Statistical Analysis

We compared first versus last criteria versions across refined items (n=8) using bootstrap methods with 10,000 iterations. Bootstrap provides robust p-values without requiring distributional assumptions that may not hold for our metrics. For non-refined items (n=9), we report performance metrics for the single iteration only. We used one-tailed tests for all metrics: expecting increases for MCC, κ_C, and accuracy, and expecting a decrease for FPR.

3.1.4 Results

RQ1: Creator Engagement in Iterative Refinement Creators used iteration effectively. Items underwent a median of 3 versions (mean 5.2 versions per item), with 58.8% of items being revised at least once (10 of 17 items). Among all 17 items, 8 (47%) had meaningful criteria version changes that we analyze as "refined items," while 9 (53%) maintained consistent criteria across runs ("non-refined items"). This iteration pattern suggests creators found a productive balance between refinement and effort: enough iteration to improve performance without excessive revision cycles.

The platform enabled rapid development. Items were developed over a median of 1 day (range: 1-4 days), addressing the time bottleneck.

Criteria became more detailed through iteration, with a median increase of 5 words from first to last version (60 to 65 words, representing an 8.3% increase). The number of individual criteria also increased modestly from an average of 1.9 to 2.2.

RQ2: Performance Improvements Through Iteration For the 8 items that underwent criteria refinement (47% of the dataset), comparing first versus last criteria versions revealed statistically significant improvements across key performance metrics (Table 1). Our primary metric, MCC, improved from 0.659 to 0.863 (p < 0.001), representing a +0.203 improvement in scoring reliability. This improvement means refined items moved from moderate to strong reliability. κ_C also improved significantly, from 0.620 to 0.860 (p < 0.001), a gain of +0.240. According to Landis and Koch (1977), this represents improvement from substantial agreement (0.61-0.80) to almost perfect agreement ($\geq$ 0.81).

False positive rates decreased from 0.148 to 0.087 (-0.060, p = 0.163). While not statistically significant in aggregate, individual items showed varied patterns. Some items achieved large FPR reductions (one item improved by 0.420). Others experienced FPR increases while creators prioritized our primary MCC metric (another item's FPR increased from 0.020 to 0.140 while its MCC improved by 0.323). Overall accuracy improved significantly from 0.818 to 0.938 (p < 0.001), representing a +0.119 improvement. Notably, 100% of refined items showed improvements in both MCC and κ_C, demonstrating that the iterative refinement process consistently led to better scoring reliability.

For example, item A-CED.A.3 asks students to interpret inequality solutions in real-world contexts. Through iteration, creators refined the criteria for greater precision. The refined criteria specified that students must **explicitly** state why a whole number is needed for the real-world scenario and **explicitly** explain why rounding down is necessary to satisfy the inequality. These refinements, which instructed the AI Scorer not to accept implied reasoning, improved the AI Scorer's MCC from 0.554 to 0.971 (Figure 4; see Appendix A for complete criteria text).

RQ3: Achievement of Reliability Standards With iterative refinement, 100% of items achieved our primary reliability standard (MCC $\geq$ 0.80), compared to only 58.8% based on first-attempt performance. This 100% success rate shows how CDP's guided refinement process makes reliable assessment development accessible to creators regardless of their psychometric expertise. This improvement suggests that CDP rescued 7 items that would have required abandonment or costly pilot-based revision. When considering both MCC and our secondary FPR threshold ($\leq$ 0.10), 76% of items (13 of 17) met both standards after CDP refinement.

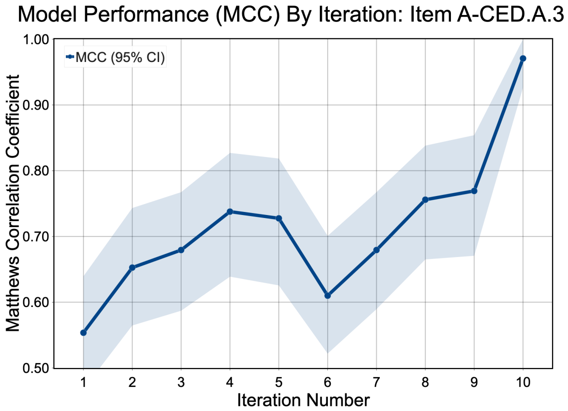


Figure 4: Item A-CED.A.3 scoring reliability across 10 iterations. Matthews Correlation Coefficient improved from 0.554 to 0.971 (95% confidence intervals shown), demonstrating how iterative refinement strengthened the scoring criteria.

The items demonstrated two development patterns. Nine items (53%) achieved strong performance immediately, meeting the MCC threshold of 0.80 without criteria changes. These items maintained consistent criteria across all runs. The remaining 8 items (47%) underwent iterative refinement. Of these refined items, only 1 (12.5%) initially met the MCC threshold but was still refined (possibly to improve other metrics or address creator concerns). Through CDP's iterative process, all 8 refined items achieved MCC $\geq$ 0.80, with a final mean MCC of 0.863.

All 8 refined items achieved the MCC threshold, but FPR outcomes varied. Five of 8 items (62.5%) met the FPR $\leq$ 0.10 threshold after refinement. This reflects the challenge of optimizing multiple metrics simultaneously. Creators sometimes prioritize specific metrics based on their assessment goals.

These results validate CDP's solution to the twin challenges of time and expertise: all items achieved reliability standards (expertise) within days rather than months (time).

4 Limitations

Three limitations shape the interpretation of these results.

First, synthetic responses cannot capture all the ways real students think. Students use unexpected

Table 1: First vs. Last Criteria Version Performance Comparison for Refined Items (n=8)

Metric	First Run [95% CI]	Last Run [95% CI]	Change	p-value	Items Improved
MCC	0.659 [0.589-0.734]	0.863 [0.833-0.900]	+0.203	<0.001***	8 (100%)
κ_C	0.620 [0.533-0.711]	0.860 [0.830-0.898]	+0.240	<0.001***	8 (100%)
FPR	0.148 [0.033-0.288]	0.087 [0.045-0.130]	-0.060	0.163	3 (37.5%)
Accuracy	0.818 [0.769-0.867]	0.938 [0.924-0.954]	+0.119	<0.001***	8 (100%)

Note: *** p < 0.001; Bootstrap tests with 10,000 iterations, one-tailed

terminology, creative analogies, and unique error patterns that synthetic generation misses. Future work must validate with real student data.

Second, we validated CDP with mathematics items and GPT-4o. While the approach should generalize to other domains using the same EYT format, criteria optimized for GPT-4o's scoring tendencies might not transfer directly to other LLMs.

Third, CDP optimizes scoring reliability but doesn't evaluate conversation quality. Criteria both evaluate and trigger follow-ups. We measured scoring, not dialogue quality. Future work should examine whether improvements in scoring reliability correlate with better conversation flow and more effective probing of student understanding.

These limitations point to clear next steps: validating with real student data, testing beyond math and GPT-4o, and measuring conversation quality.

5 Conclusions

In order to create effective conversation-based assessments, we need effective criteria for scoring them. These criteria are traditionally difficult and time-consuming to develop. The Criteria Development Platform addresses this challenge through pre-pilot optimization with synthetic data. Our analysis of 68 development cycles across 17 mathematics items demonstrates CDP's impact: success rates improved from 59% to 100%, rescuing 7 items from abandonment or costly pilot revision. The eight items that underwent refinement showed substantial gains, with MCC improving from 0.659 to 0.863. CDP solves both traditional CBA development challenges: reducing timelines from months to days (median 1 day) while enabling non-technical experts to achieve reliable results through guided refinement.

These results have broader implications for educational technology. Pre-pilot optimization with synthetic data provides an effective approach when authentic data is expensive or unavailable. The platform's transparency shows creators exactly why scoring succeeds or fails, transforming development from intuition to evidence. By making reliable assessment development accessible to educators without specialized expertise, tools like CDP enable more practitioners to create LLM-based assessments that measure deep understanding.

6 Appendix A: Example Criteria Changes

This appendix documents the criteria refinement process for Item A-CED.A.3, which improved from MCC = 0.554 to 0.971.

Students must explain two things: why decimal solutions need whole number rounding, and why rounding down (not up) satisfies the constraint.

6.1 First Criteria Version (MCC = 0.554)

The initial criteria were:

- **Criterion 1:** Student recognizes that the answer has to be a whole number of rides in order to make sense in the real world.

- **Criterion 2:** Student acknowledges that rounding the decimal answer down to the lower whole number is necessary because rounding up to the higher whole number makes the inequality that defines the number of credits no longer true.

6.2 Final Criteria Version (MCC = 0.971)

Testing revealed the AI accepted implied reasoning when explicit statements were needed. Revised:

- **Criterion 1:** Student must explicitly state reasoning for rounding to a whole number that includes making sense in the real-world (for example, "it does not make real-world sense for a quantity of rides to be a fraction or decimal"). It is not correct for a student to imply reasoning or to only say that they rounded down.

- **Criterion 2:** Student acknowledges that rounding the decimal answer down to the lower whole number is necessary to satisfy the inequality. Student must explicitly refer to the inequality or explain why they round down in the context of the problem (example: the most number of rides without going over in credits). It is not correct for a student to imply reasoning.

References

Jan Bergerhoff, Johannes Bendler, Stefan Stefanov, Enrico Cavinato, Leonard Esser, Tommy Tran, and Aki Härmä. 2024. Automatic conversational assessment using large language model technology. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, pages 39–45, Porto Vlaams-Brabant Portugal. ACM.

Scott Frohn, Tyler Burleigh, and Jing Chen. 2025. Automated Scoring of Short Answer Questions with Large Language Models: Impacts of Model, Item, and Rubric Design. In *Artificial Intelligence in Education*, volume VI of *Lecture Notes in Artificial Intelligence*, pages 44–51, Palermo, Italy. Springer.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 300–304, Atlanta GA USA. ACM.

G. Tanner Jackson, Katherine E. Castellano, Debra Brockway, and Blair Lehman. 2018. Improving the Measurement of Cognitive Skills Through Automated Conversations. *Journal of Research on Technology in Education*, 50(3):226–240.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159.

OpenAI. 2024. Introducing ChatGPT.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk. ACM.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Scott Wood, Erin Yao, Lisa Haisfield, and Susan Lottridge. 2021. Establishing Standards of Best Practice in Automated Scoring. Technical report, ACT, Inc. Publication Title: ACT, Inc. ERIC Number: ED616491.

Seyma N. Yildirim-Erbasli and Okan Bulut. 2023. Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment. *Computers and Education: Artificial Intelligence*, 4:100135.

Diego Zapata-Rivera, Tanner Jackson, and Irvin R. Katz. 2015. Authoring Conversation-based Assessment Scenarios. In Robert A. Sottilare, Arthur C. Graesser, Xiangen Hu, and Keith Brawner, editors, *Design Recommendations for Intelligent Tutoring Systems, Volume 3: Authoring Tools & Expert Modeling Techniques*, pages 169–178. U.S. Army Research Laboratory, Orlando.