Within One Point of TEA Raters in 98.5% of Cases: The Design, Validity, and Limits of CoGrader's Automated STAAR Constructed Response Scoring

Whitepaper (Short) · June 2026

A technical appendix with full statistical methods is available on request.

Executive Summary

The Formative Assessment Gap in Texas Districts

The Texas Education Agency's (TEA) 2023 STAAR redesign expanded constructed-response items sharply, producing six to seven times more written items per test (Texas Tribune, 2024). To score the operational exam at that volume, the state moved to a hybrid model pairing an automated scoring engine with human raters. Year-round practice writing was left where it was. Teachers still hand-score every practice essay, benchmark, and interim, at a reported 15 to 20 minutes each. Districts are left choosing between assigning less writing and returning feedback weeks later, after the revision window has closed.

The Solution: CoGrader’s Automated STAAR Scoring System

CoGrader brings semi-automated, TEA-aligned scoring into the classroom. The CoGrader system evaluates student practice writing against the state's published rubrics and returns a score with rubric-aligned feedback in roughly three seconds. It does not replace the educator. A human-in-the-loop protocol is in place with the teacher, who reviews it, adjusts where needed, and holds final authority in delivering high quality feedback and a suggested score directly to the learner.

Empirical Validation and Technical Fidelity

The CoGrader research team benchmarked the system against TEA's own official scores using a public corpus of 600 scored responses spanning eight scoring modules, four item types, three grade bands, and two languages. These were the results:

- **98.5% accuracy within one point of TEA raters,** meeting the state's operational agreement standard.
- **79.3% exact match with the official TEA score; 72.4% on extended responses alone. Trained human rater pairs on complex constructed responses typically fall in the 60-75% range.**
- **Roughly three seconds per response, to deliver comprehensive, rubric-aligned feedback compared to 15 to 20 minutes of manual grading by TEA raters.**

On essays scored and published by TEA, CoGrader came within one point of the official score 98.5% of the time.

That is the same bar Texas uses when two of its own trained human raters score the same essay. CoGrader scoring is precise enough to guide practice and benchmark writing during the school year. CoGrader provides Texas curriculum leaders and campus principals with a reliable, scalable, and reproducible solution to elevate student writing proficiency and ensure STAAR readiness before test day.

Because a CoGrader score returns in approximately 3 seconds rather than 15 to 20 minutes of teacher time, constructed response practice becomes possible at the volume students need. Teachers can assign more writing because feedback is handled, students receive rubric-aligned feedback while revision is still possible, and teachers and district leaders can see student writing performance against STAAR rubrics during the year, when instruction can still respond, rather than after it ends.

Introduction

CoGrader's automated STAAR scoring system landed within one point of TEA human raters in 98.5% of cases across 600 officially scored responses spanning eight scoring modules. It matched the TEA score exactly in 79.3% of cases, at roughly three seconds per response. This paper describes the system's design, the validation evidence behind those figures, and the limits of what they support.

Background: Texas Districts Face No-Win Compromise

The 2023 STAAR redesign capped multiple-choice items at no more than 75% of total points and added new item types, including a heavy emphasis on constructed response (TEA, n.d.). The result was six to seven times more written items per test than under the previous design (Texas Tribune, 2024). To score that volume on the operational exam, TEA adopted a hybrid model in which an automated engine scores every constructed response and roughly a quarter

are routed to human raters for rescoring, along with any response flagged for low confidence (TEA & Cambium Assessment, 2023; Texas Tribune, 2024).

The operational test is scored automatically. Everything leading up to it is not. The labor-intensive work of scoring practice essays, benchmarks, and interim assessments falls on individual teachers, and scoring a single extended response against the STAAR rubric takes 15 to 20 minutes. The volume of practice writing students need over a school year exceeds the hours teachers have. Districts face a dilemma with no good side: assign less writing, or make students wait weeks for feedback and miss the window in which revision improves learning.

Bridging the Formative Assessment Gap

CoGrader offers a solution to bridge this diagnostic gap with its STAAR scoring system that brings automated, TEA-aligned scoring directly into the classroom during the instructional year. CoGrader's AI-powered evaluation assistant reviews student practice writing against exact, state-aligned rubrics, returning timely and specific feedback in approximately 3 seconds rather than 15 to 20 minutes of teacher time, allowing constructed response practice to become possible at the volume students need.

This rapid turnaround of rubric-aligned feedback allows educators to confidently assign more writing and adjust daily instructions instantly, enabling students to revise their work while the concepts are fresh in their minds.Now district leaders can measure student writing performance against STAAR rubrics during the year, when teacher instruction can respond, not after it ends.

For automated practice feedback to be instructional, it must be highly accurate. This paper raises and answers a critical question for Texas school administrators: Does CoGrader score student writing the same way official TEA raters score it?

Evaluating the Fidelity of Automated Scoring

To evaluate this fidelity, a validation study was conducted across 600 officially scored TEA essays spanning eight distinct writing item types. This paper outlines the architecture of CoGrader's automated system and presents data validating its alignment with Texas' state testing standard.

Automated scoring of constructed responses is not itself new to Texas. TEA scores the operational STAAR with an automated engine working alongside human raters, reducing its temporary scorer hiring from roughly 6,000 to under 2,000 in the first year of the hybrid model (Texas Tribune, 2024). The state's own adoption establishes that automated scoring already operates inside Texas assessment, which reframes the district question from whether to use it to which systems demonstrate fidelity to how the state actually scores. AI-assisted grading is also an active research area more broadly, with recent work including a human-LLM collaborative system for grading project reports, independently also named CoGrader (Chen et al., 2025).

The district side of the equation remains unaddressed, as no equivalent feedback capability exists for the formative work that prepares students for these exams. The question for districts is which systems can fill that gap with demonstrated fidelity to how the state actually scores.

The following analysis comes from the CoGrader research team. CoGrader has scored more than one million student essays across 1,000+ schools, has backing from UC Berkeley's SkyDeck, and is the subject of an active U.S. Department of Education IES funded study focused on accelerating student writing proficiency. While independent peer-reviewed work has examined instructor experience with the tool in university writing assessment (Alsalem, 2024), this paper focuses solely on scoring accuracy.

System Construct and Design

The CoGrader system combines a current-generation production large language model with the following design elements:

- Calibration to TEA's own standard: each module uses the publicly released TEA STAAR scoring guides as in-context training material,
- A task-specific module architecture, one module per STAAR item type and grade band, each with its own calibration logic, and
- A human-in-the-loop protocol in which the teacher reviews each score, adjusts where needed, and retains final authority.

What the system scores

STAAR does not use one writing rubric. Each item type is scored on its own scale, and the same item type is scored differently across grade bands. So the system is not one general scorer. It is eight task-specific modules, each calibrated against the TEA scoring guide for that exact item, grade band, and language. The following eight modules make up the system as described in the table below. Six modules cover Extended Constructed Response (ECR) and two cover Short Constructed Response (SCR). ECR coverage spans Grades 3-5, 6-8, and English I/II in English, plus Grades 3-5 in Spanish. Argumentative coverage is currently partial, with modules at Grades 6-8 English and Grades 3-5 Spanish. Additional modules are in development.

Module	Grade band	Language	Rubric scale
ECR Informational	3-5	English	0-5
ECR Argumentative	3-5	Spanish	0-5
ECR Informational	3-5	Spanish	0-5

ECR Informational	6-8	English	0-5
ECR Argumentative	6-8	English	0-5
ECR Informational	HS (English I/II)	English	0-5
Short Response Reading	All	English	0-1-2
Short Response Writing/Revising/Editing	All	English	0-1

How the system is built

A fair question about any automated scorer is where its judgment comes from. Some scoring systems learn from a private collection of student essays. When that happens, a district has no way to check what the system learned, or whether the standard it absorbed matches the one Texas actually applies.

CoGrader does not work that way. No custom scoring model was trained for this work. Each of CoGrader's eight modules reads TEA's publicly released scoring guides directly and uses them as calibration material. The standard CoGrader applies is the state's published standard, and any teacher or administrator can read the same guides CoGrader reads.

That leaves a fair question about what stays proprietary. CoGrader's scoring engine identity, configuration files, and calibration logic are documented internally and shared with district auditors under a confidentiality agreement.

Scoring Validity

The CoGrader system was benchmarked on 600 essays drawn from publicly released TEA scoring guides and TEA-approved rater training material (TEA, 2022). These scores constitute the formal gold standard TEA uses to train its own human raters.

Pooled across the 8 modules, the system's score landed within 1 point of the TEA grader's score in 98.5% of cases, and matched it exactly in 79.3% of cases. The within-one-point threshold is not a benchmark we chose. It follows TEA's own rule for human ECR scoring. Under STAAR's scoring protocol, each human-scored ECR is read by two trained scorers, and the two scores are added. On each rubric trait, scores that are one point apart (adjacent) both stand. Scores that are further apart go to a third resolution read. For SCRs, TEA requires an exact match between scorers (TEA & Cambium Assessment, 2023, Table 3, pp. 8-9). So for SCR items, exact agreement is the fair comparison.

On STAAR ECRs in spring 2023, two trained TEA raters gave the exact same score on a rubric trait about 63% to 75% of the time, depending on grade and trait (TEA & Cambium Assessment, 2023, Table B7, pp. 37-38). NAEP treats 60% exact agreement as acceptable for a complex six-point writing item (National Center for Education Statistics, n.d.-a).

The relevant question for an automated scorer, then, is not whether it matches the official score in every case; trained human raters do not. It is whether the system agrees with a trained rater about as consistently as a second trained rater would. On this benchmark it does. The system meets the agreement standard in 98.5% of cases pooled, and 96.8% on extended response alone. Its exact-match rate is 79.3% pooled and 72.4% on extended response. On NAEP's six-point holistic writing scale, two trained raters agree exactly 57% to 67% of the time. CoGrader's rate sits above that range. For CoGrader, our belief is that a practice score is only instructionally valid to the extent it predicts how the state will score the same writing.

Module	Number of Items	Exact match	Within 1 point
SCR - Writing	166	89.8%	100.0%
SCR - Reading	151	80.8%	100.0%
ECR	283	72.4%	96.8%

ECR results follow:

Module	Number of Essays	Exact match	Within 1 point
ECR 3-5 Informational (English)	72	84.7%	98.6%
ECR 6-8 Informational	69	75.4%	97.1%
ECR 6-8 Argumentative	15	73.3%*	100.0%
ECR HS Informational	61	70.5%	95.1%
ECR 3-5 Argumentative (Spanish)	12	66.7%*	91.7%
ECR 3-5 Informational (Spanish)	54	55.6%	96.3%

* Small samples; indicative only. Confidence intervals, reliability coefficients, and full methods appear in the technical appendix.

Two considerations qualify these results and are described as follows.

1. All 600 responses are TEA scoring examples and rater training material, the cleanest illustrations of each score point. Results should be read as a ceiling relative to operational classroom writing, which is messier.
2. The same public scoring guides supply both the in-context calibration material and the validation corpus. We separated them mechanically. Fidelity on responses drawn from outside that material has not yet been measured.

Commitment to Continuous Improvement

Adoption follows the phased structure districts already use for scoring changes: evidence review, a scoped trial, then implementation decisions. Before any agreement is signed, a live demonstration is run on district-selected essays, a blind agreement test is conducted on essays scored by the district's own raters, and module-by-module fit, including the limitations above, is reviewed with the district's assessment team alongside operational requirements (integration, FERPA review, audit logging, arbitration workflow, drift monitoring).

Districts have good reason to scrutinize any scoring instrument that reaches student writing. The approach here is to earn that scrutiny: publish the design, validate against the state's own gold standard, disclose limitations before others find them, and keep the teacher in final authority over every score. Districts interested in a pilot can request a 30-minute discovery call through their existing point of contact or CoGrader's website.

References

- Alsalem, M. S. (2024). EFL teachers' perceptions of the use of an AI grading tool (CoGrader) in English writing assessment at Saudi universities: An activity theory perspective. *Cogent Education, 11*(1), Article 2430865. https://doi.org/10.1080/2331186X.2024.2430865
- Cambium Assessment. (n.d.-a). *Communicating to the public about machine scoring* [White paper]. https://files.portal.cambiumast.com/corporate-site/documents/CAI-Cambium-CommunicatingPublicMachineScoring-WhitePaper.pdf
- Cambium Assessment. (n.d.-b). *Smarter Balanced interim assessments: Automated scoring FAQ*. https://smarterbalanced.alohahsap.org/content/contentresources/en/Smarter_Interim_Automated_Scoring_FAQ.pdf
- Chen, Z., Wang, J., Li, Y., Li, H., Shi, C., Zhang, R., & Qu, H. (2025). *CoGrader: Transforming instructors' assessment of project reports through collaborative LLM integration* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.20655
- National Center for Education Statistics. (n.d.-a). *NAEP technical documentation: Constructed-response interrater reliability*. https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.aspx

- National Center for Education Statistics. (n.d.-b). *NAEP technical documentation: Scoring*. https://nces.ed.gov/nationsreportcard/tdw/scoring/
- Texas Education Agency. (n.d.). *STAAR redesign*. https://tea.texas.gov/student-assessment/assessment-initiatives/staar-redesign
- Texas Education Agency. (2022, October 20). *STAAR redesign constructed response resources and updated calendar of events* [To the Administrator Addressed letter]. https://tea.texas.gov/about-tea/news-and-multimedia/correspondence/taa-letters/staar-redesign-constructed-response-resources-and-updated-calendar-of-events
- Texas Education Agency & Cambium Assessment. (2023). *STAAR hybrid scoring study, methods and results: Spring 2023 items*. https://tea.texas.gov/data-reports/reports-and-studies/2023-staar-hybrid-scoring-study-2.pdf
- Texas Tribune. (2024, April 9). *How Texas will use AI to grade this year's STAAR tests*. https://www.texastribune.org/2024/04/09/staar-artificial-intelligence-computer-grading-texas/