Within One Point of TEA Raters in 98.3% of Cases: A 1,764-Response Benchmark of CoGrader's Automated STAAR Constructed-Response Scoring

CoGrader Research · October 2026

Executive Summary

CoGrader's grading engine landed within one score point of official Texas Education Agency (TEA) raters on 98.3% of 1,764 redesigned STAAR constructed responses (and 99.5% across all 2,437 scored rubric traits). Additionally, CoGrader's grading engine was able to match the TEA-assigned rating for the exact number of points on 80.4% of the 1,764 responses across 154 TEA prompts.

This study expands CoGrader's initial 600-response STAAR validation nearly threefold, evaluating 1,764 student responses drawn from released TEA scoring guides and rater-training materials across grades 3 to 8, English I, English II, Biology, and U.S. History. Due to the fact that the redesigned STAAR has two distinct constructed-response item types with different scoring scales and operational rules, we will present results both separately and combined:

- Separated results for **Short Constructed Responses (SCR) (scale of 0 to 2, 1,091 total SCR responses from 106 prompts):** CoGrader was able to match the official TEA rating on 90.5% of all responses and fall within one point of the TEA rating 99.7% of the time (QWK = 0.92). CoGrader met TEA's exact-match operational standard across Reading Language Arts, Science, and Social Studies.
- Separated results for **Extended Constructed Responses (ECR) (composite scale of 0 to 5 for two traits across all responses, 673 total ECRs from 48 prompts):** CoGrader was able to match the official TEA rating on 75.0% of all traits and fall within one point of the TEA rating on 99.4% of all traits (96.0% fell within one point on the 0 to 5 composite essay score, QWK = 0.91). Two trained TEA human raters agree on an exact trait rating on approximately 63% to 75% of all operational STAAR ECRs and accept adjacent scores within one point.
- **High-Confidence Routing:** In alignment with TEA's hybrid scoring model, which allows high-confidence automated ratings to "stand" and borderline ratings to be flagged for human review, CoGrader achieved exact agreement at 97.8% on SCRs and 85.7% per-trait exact agreement (99.1% within one point per trait) on ECRs among the top half of responses by confidence. It rose to exact agreement at 99.4% on SCRs and 94.3% per-trait exact agreement (100% within one point per trait) on ECRs among the top 30%.

Measure	Result
Total number of STAAR responses evaluated	1,764 officially scored responses across 154 prompts (2,437 scored rubric traits)
Percentage of all responses that were within one score point of the TEA rating	98.3% (99.5% across all 2,437 scored rubric traits)
Percentage of all responses that were an exact match for the official TEA rating	80.4% (81.9% across all 2,437 scored rubric traits)
Percentage of Short Constructed Responses (SCR, 0 to 2 scale)	90.5% exact match, 99.7% within one point, 0.92 QWK (1,091 responses across 106 prompts)

Measure	Result
Percentage of Extended Constructed Responses (ECR, per trait)	75.0% exact match per trait, 99.4% within one point per trait (TEA human rater pair: 63% to 75%)
Percentage of Extended Constructed Responses (ECR, 0 to 5 total)	96.0% within one point, 64.2% exact composite match, 0.91 QWK (673 essays across 48 prompts)
High-confidence tier (top 50% by confidence)	97.8% exact on SCRs; 85.7% per-trait exact and 99.1% within one point on ECRs

Because CoGrader returns a rubric-aligned score and actionable written feedback in roughly three seconds rather than 10 to 20 minutes of teacher grading time, Texas districts can assign constructed-response practice at the volume students need while keeping teachers in final authority over every grade.

Why this matters: The formative assessment gap in Texas

The Texas Education Agency (TEA) changed how it assessed students as part of the 2023 STAAR redesign. TEA limited the amount of multiple-choice questions on each subject test to no more than 75 percent of total test points. In addition, the new tests include written constructed-response items across grade levels and subjects, resulting in six to seven times more student writing per test compared to the prior design (TEA, n.d.; Texas Tribune, 2024). Therefore, in order to assess that volume of student writing on the spring operational exam, TEA adopted a hybrid scoring model in which a machine-based engine assesses the constructed responses and sends approximately one-quarter of the responses, as well as lower-confidence responses, to human raters (TEA & Cambium Assessment, 2023).

Although the spring operational exam is assessed by machines, classroom and benchmark writing throughout the school year are not. While full-scope STAAR Interim Assessments do not contain constructed-response items that require manual scoring, teachers can add SCR and ECR prompts using the Texas Formative Assessment Resource (TFAR), but TFAR does not provide automatic scoring of the responses to the prompts added by teachers. Thus, individual teachers must manually score every practice essay, benchmark, and short answer. With each extended response taking 10 to 20 minutes to grade, a middle or high school teacher with 130 students must spend over 25 hours grading just one round of benchmarks. That bottleneck creates an educational dilemma for districts: either reduce the amount of writing assigned to students below what the redesigned STAAR demands, or delay providing feedback to students until weeks after they have completed assignments, which would be too late to allow students to revise assignments and educators to adjust instruction.

CoGrader addresses that gap by incorporating TEA-aligned automated scoring and rubric-based feedback into the classroom instructional process during the instructional year through a human-in-the-loop process in which the teacher reviews, adjusts if needed, and approves all scores.

The benchmark: Separating ECR and SCR across 1,764 responses

There are two major components of STAAR that need to be addressed when rigorously evaluating the automated scoring of STAAR: the ways in which TEA constructs and grades its two constructed-response format types:

- **Short Constructed Responses (SCR, 0 to 2 points):** Short written responses (usually 40 to 100 words) that assess reading comprehension using text evidence, writing, revising, and editing skills, science reasoning, or social studies content accuracy. For operational STAAR scoring, TEA requires **complete agreement** between two raters on SCR items (TEA & Cambium Assessment, 2023, Table 3).
- **Extended Constructed Responses (ECR, 0 to 5 points across two traits):** Multi-paragraph informational or argumentative essays written in response to a reading passage. TEA scores ECRs across two separate analytic traits: **Organization and Development of Ideas (0 to 3 points)** and **Conventions (0 to 2 points)**, where a score of 0 on Organization and Development automatically results in 0 on Conventions.

In operational human scoring, two independent raters score each trait when humans score ECRs. If there is adjacent agreement on a particular trait (i.e., a discrepancy of no more than one point on a particular trait), then both ratings remain. Otherwise, the item goes to a third reader to resolve the discrepancy. In the earlier June 2026 study of 600 responses, SCR and ECR items were reported primarily as a combined pool. In the expanded research of 1,764 officially scored redesigned STAAR responses to 154 prompts, we report both a combined benchmark and separate full-scale assessments of the 1,091 SCR responses and 673 ECR essays across grades 3 to 5, 6 to 8, and high school (English I, English II, Biology, and U.S. History).

Method

Each student response within the benchmark has been assigned an official score by TEA raters in published STAAR scoring guides and rater-training documents for STAAR testing from 2022 through 2026.

Evaluation Protocol. CoGrader's grading engine graded each student response based on the assignment prompt, the accompanying reading passage or stimulus, and TEA's published rubrics and scoring standards for STAAR, with no knowledge of the student's official TEA score. To ensure that our findings are due to the ability of CoGrader to generalize to new assignments rather than simply remembering specific prompts during calibration, the calibration process was restricted to an out-of-fold-by-prompt approach: when scoring students' responses to any particular STAAR prompt, all responses associated with that same prompt and reading passage were excluded from the calibration set. An additional, pre-declared, one-time-only evaluation was conducted on a sealed holdout group of prompts (19 ECR prompts and 42 SCR prompts) to evaluate whether CoGrader would maintain its performance on prompts excluded from calibration.

Metrics Reported. Following guidelines established by the field of educational measurement and TEA's hybrid scoring technical reports (TEA & Cambium Assessment, 2023), we report three key metrics:

- **Exact Match:** The percentage of responses (or rubric traits) for which CoGrader's score matches the official TEA score. As noted above, exact match is the metric used by TEA as its primary operational criterion for evaluating SCRs and comparing ECRs by trait.
- **Within One Score Point (adjacent agreement):** The percentage of responses (or traits) for which CoGrader's score falls within one point of the official TEA rating. This corresponds with TEA's operational non-resolution policy for human ECR rater scores.

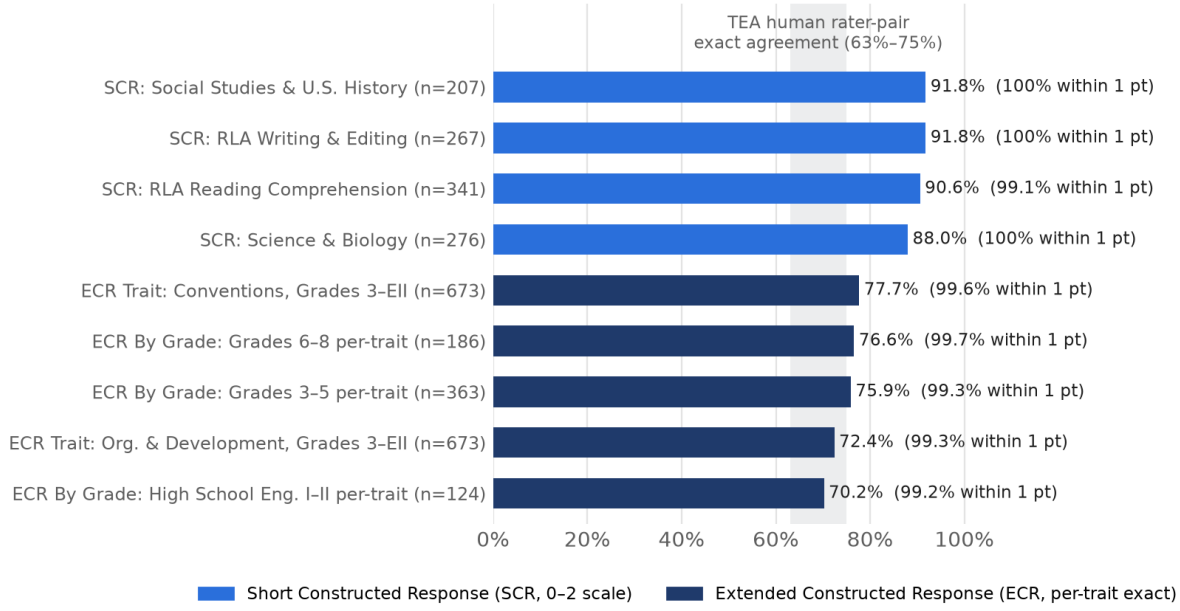
- **Quadratic Weighted Kappa (QWK):** The standard measure of chance-corrected reliability for an educational assessment. It measures how well raters agree on test ratings and ranges from .00 (chance agreement) to 1.00 (complete agreement). Quadratic weighted kappa penalizes larger point differences quadratically.

Results

Across all 1,764 redesigned STAAR responses (154 prompts), CoGrader landed within one point of the official TEA score in 98.3% of cases and matched the official score exactly in 80.4% of cases. Evaluated across all 2,437 individual rubric-trait scores (1,091 SCR scores plus 1,346 ECR trait scores), CoGrader achieved 81.9% exact match and 99.5% agreement within one point.

Assessment category	Prompts	Evaluated units	Scale	Exact match	Within 1 point	QWK
Short Constructed Response (SCR)	106	1,091	0 to 2 (1 trait)	90.5%	99.7%	0.92
Extended Constructed Response (ECR, trait level)	48	1,346 traits (673 essays)	0 to 3 (OD), 0 to 2 (CV)	75.0%	99.4%	0.87 (OD) / 0.82 (CV)
Extended Constructed Response (ECR, composite)	48	673	0 to 5 (sum of 2 traits)	64.2%	96.0%	0.91
All STAAR responses (response level)	**154**	**1,764**	**0 to 2 and 0 to 5**	**80.4%**	**98.3%**	**0.91 to 0.92**
All STAAR evaluations (trait level)	**154**	**2,437 traits**	**All traits**	**81.9%**	**99.5%**	**0.91 median**

Figure 1: CoGrader matched or exceeded TEA's human rater-pair agreement band for every ECR rubric trait and every SCR subject.



Exact agreement with official TEA scores across all 1,091 Short Constructed Responses and 673 Extended Constructed Responses (1,346 scored traits). The shaded band represents the range of exact agreement between human rater pairs on operational Spring 2023 ECR traits (63% to 75%; TEA & Cambium Assessment, 2023, Table B7).

Short Constructed Responses (1,091 responses across 106 prompts). On SCR items, which require exact agreement between raters according to TEA's policy, CoGrader exactly matched the official TEA score 90.5% of the time. Additionally, it agreed with the official TEA score within one point on SCR items approximately 99.7% of the time and had a QWK of .92. Furthermore, CoGrader performed well at distinguishing incorrect or unsubstantiated answers from those providing some evidence of learning at the critical zero-versus-credit boundary, achieving an accuracy of 94.7%. CoGrader also showed consistently high performance across all four tested content areas:

SCR subject module	Responses	Exact match	Within 1 point	QWK	0-vs-credit accuracy
Social Studies and U.S. History	207	91.8%	100.0%	0.94	97.6%
RLA Writing, Revising, and Editing	267	91.8%	100.0%	0.84	92.1%
RLA Reading Comprehension	341	90.6%	99.1%	0.91	95.0%
Science and Biology	276	88.0%	100.0%	0.91	94.6%
All Short Constructed Responses	**1,091**	**90.5%**	**99.7%**	**0.92**	**94.7%**

On the Short Constructed Responses in Social Studies and U.S. History (207 responses) and RLA Writing and Editing (267 responses), CoGrader reached 91.8% exact agreement and 100.0% agreement within one point.

In RLA Reading Comprehension (341 responses), in which students have to find an implicit inference as well as the corresponding verbatim textual evidence, CoGrader scored 90.6% exact agreement and 0.91 QWK.

In Science and Biology (276 responses), CoGrader achieved an 88.0% exact match, 100.0% within one point, and a QWK of 0.91. On the pre-declared sealed holdout test (287 responses across 42 unseen prompts), SCR exact agreement was 89.9% (99.7% within one point, 0.91 QWK), confirming that accuracy holds on completely unseen prompts.

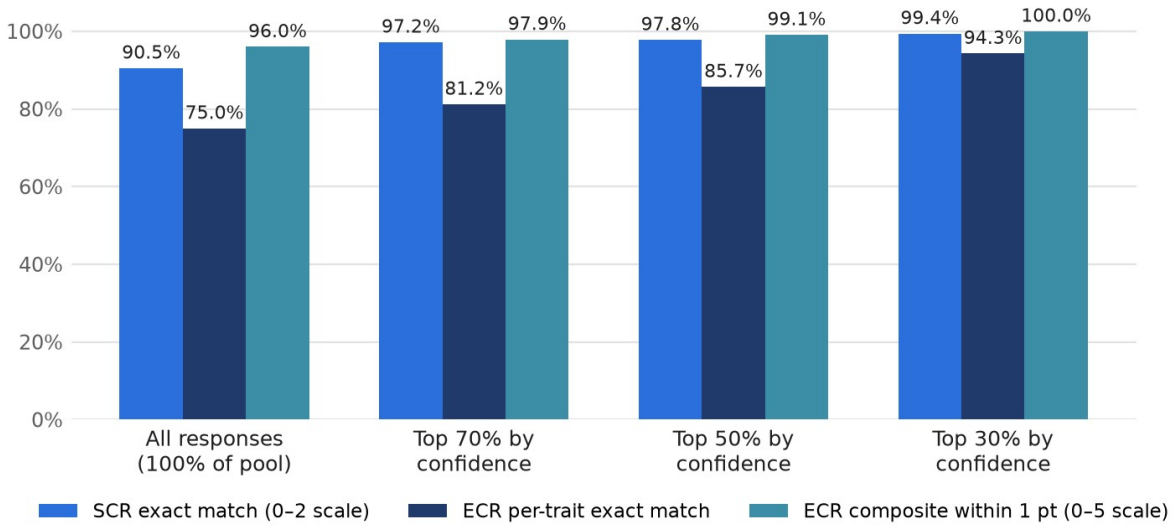
Additionally, CoGrader was evaluated on 673 extended constructed responses (ECRs) across 48 STAAR prompts. The results were compared with the performance of two trained TEA human raters on the operational spring 2023 exam, who agreed on the exact trait score 63% to 75% of the time, depending on grade and trait (TEA & Cambium Assessment, 2023, Table B7, pp. 37 to 38). Across all 673 ECR essays (1,346 trait scores), CoGrader matched the official TEA trait score 75.0% of the time and landed within one point on 99.4% of traits. The comparison showed the following:

ECR evaluation dimension	Evaluated units	Exact match	Within 1 point	QWK	Human rater benchmark
Organization and Development of Ideas (0 to 3)	673	72.4%	99.3%	0.87	63% to 75%
Conventions (0 to 2)	673	77.7%	99.6%	0.82	63% to 75%
All ECR rubric traits combined	**1,346 traits**	**75.0%**	**99.4%**	**0.85**	**63% to 75%**
ECR composite essay score (0 to 5 total)	**673 essays**	**64.2%**	**96.0%**	**0.91**	**57% to 67% (NAEP 6-pt)**

The composite essay score is the sum of the Organization and Development and Conventions trait scores. To achieve an exact match of this composite score, the sum of both traits must match the official TEA composite score. CoGrader achieved an exact match of the six-point composite score with the official TEA score on 64.2 percent of essays; landed within one point of the official TEA composite score on 96.0 percent of essays (97.0 percent on the 268-essay sealed holdout test); had a QWK of .91; and was also at the top of the NAEP 57 percent to 67 percent exact-agreement benchmark for two human raters identifying a student's score based on a six-point holistic writing scale (NCES, n.d.).

Confidence routing: Mirroring TEA's Hybrid Scoring Model. CoGrader's confidence routing mirrors TEA's hybrid rating system. TEA uses its automated machine-scoring platform to evaluate each student response. Based on this evaluation, TEA sends lower-confidence student responses to be evaluated by human raters. CoGrader provides a calibration-based confidence score for each response to allow educators to use their evaluation time to assess borderline responses while quickly verifying high-confidence assessments.

Figure 2. Responses with high confidence produce 97.8% to 99.4% exact matches on SCRs and 99.1% to 100% agreement within one point on ECRs.



The accuracy of CoGrader compared with the official TEA scores across all confidence levels, using data from 1,091 SCRs and 673 ECR essays.

Confidence tier	SCR n	SCR exact	SCR within 1	ECR n	ECR trait exact	ECR total within 1	ECR QWK
All responses (100% of pool)	1,091	90.5%	99.7%	673	75.0%	96.0%	0.91
Top 70% by confidence	763	97.2%	100.0%	471	81.2%	97.9%	0.95
Top 50% by confidence	545	97.8%	100.0%	336	85.7%	99.1%	0.97
Top 30% by confidence	327	99.4%	100.0%	201	94.3%	100.0%	0.99

On the 50% of essays where CoGrader is most confident, per-trait exact agreement on ECRs rises to 85.7% (80.7% exact on the 0 to 5 composite score and 99.1% within one point, with 0.97 QWK), while SCR exact agreement reaches 97.8%. On the top 30% by confidence, SCR exact agreement

reaches 99.4%, and ECR composite agreement reaches 92.5% exact and 100.0% within one point (0.99 QWK).

Classroom Implementation and District Validation

For Texas curriculum leaders and campus principals, these results demonstrate that automated classroom practice scoring can match the reliability of the state's own human rater pool while eliminating the multi-week grading bottleneck:

- **Teacher in the Loop, Always.** CoGrader does not replace the educator. Every score and rubric comment is presented as a draft for teacher review, with confidence indicators that highlight borderline papers where a teacher's closer look adds the most value.
- **Objective, Fatigue-Free Consistency.** Grading 120 extended responses by hand introduces inevitable rater fatigue and halo effects between the first paper of the evening and the fortieth. CoGrader applies the exact same STAAR rubric standard to every draft, giving teachers an objective second read and allowing students to revise while the lesson is still active.
- **Phased district validation.** District adoption follows a structured verification protocol: an initial evidence review, a live demonstration on district-selected writing samples, and a blind agreement check against essays scored by the district's own curriculum team, alongside FERPA, integration, and security reviews.

Appendix: Benchmark Breakdown by Grade Band and References

Table A1 summarizes performance across all three STAAR grade bands (Grades 3 to 5, Grades 6 to 8, and high school EOC) for both Extended and Short Constructed Responses. All 1,764 responses come from public TEA STAAR Scoring Guides and Rater-Training Materials; no private student data from CoGrader customer districts was used in this benchmark.

Table A1. STAAR benchmark agreement by grade band and item type (1,764 responses across 154 prompts)

Grade band	Item type & scale	Prompts	Responses	Exact match	Within 1 point	QWK
Grades 3 to 5	ECR (0 to 5, 2 traits)	26	363	75.9% trait / 65.6% total	99.3% trait / 96.1% total	0.92
Grades 6 to 8	ECR (0 to 5, 2 traits)	13	186	76.6% trait / 65.6% total	99.7% trait / 97.3% total	0.92
High School (English I and II)	ECR (0 to 5, 2 traits)	9	124	70.2% trait / 58.1% total	99.2% trait / 93.5% total	0.88
Grades 3 to 5	SCR (0 to 2)	29	278	88.8%	100.0%	0.91
Grades 6 to 8	SCR (0 to 2)	42	434	90.6%	100.0%	0.92
High School (Eng. I and II, Biology, U.S. History)	SCR (0 to 2)	36	379	91.6%	99.2%	0.92

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

- National Center for Education Statistics. (n.d.). *NAEP technical documentation: Constructed-response interrater reliability*. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.aspx
- 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/