STAAR Writing ASE Misconceptions & Facts

REGION 13
EDUCATION SERVICE CENTER

Welcome to ESC Region 13's guide on the STAAR Automated Scoring Engine (ASE) for Constructed Response Scoring. Here, we address common misconceptions and provide clear insights into the scoring process. This chart compares common misconceptions with facts to help you better understand how the ASE really works in scoring STAAR constructed responses.

MISCONCEPTIONS	FACTS
Automated Scoring Engine Technology History	
"Automated Scoring Engine technology is new and untested." "We have never used this technology before."	Automated Scoring Engine technology is over a decade old and is widely used, including in Texas.* • The technology for the ASE has been around for 10+ years. • The Texas Success Initiative Assessment (TSIA) (used annually by 180K+ Texas students to graduate) uses an ASE for scoring. • 21+ states (including CA, FL, CO, UT, and VA) use an ASE to help score their annual state assessments. • The GRE, GED, GMAT, and TOEFL use ASEs for scoring.
Automated Scoring Engine Process	

"The Automated Scoring Engine (ASE) is the sole determinant of ECR scores."	Before the actual event of scoring constructed responses, every step of the preparation process is completely dependent on human input.
	• Field Testing
	  ○ All extended constructed responses are double human-scored.
	• Anchor Approval Meeting
	  ○ Humans identify anchor responses as field test examples. This illustrates the range of responses at each rubric score point.
	• Preparing Human Scorers and Automated Scoring Engine
	  ○ Human scorers are trained through the anchor responses.
	  ○ The Automated Scoring Engine is programmed through ~3,000 hand-scored field test responses and human-identified anchor responses.
	• Scoring
	  ○ Hybrid scoring: Human raters support the Automated Scoring Engine through calibration checks and scoring non-standard responses during the administration window.
	  ○ At least 25% of the responses to any constructed response item are routed to human scorers as "quality control." Any items with condition codes and low confidence are a part of this 25%.
	    — Condition codes include the following: response is blank, used too few words, is mostly duplicated text, is written in another language, consists mainly of stimulus material, has ideas that ASE does not recognize, used language that is reflective of an off-topic response.

MISCONCEPTIONS	FACTS

ASE Programming

"The Automated Scoring Engine (ASE) is totally computerized, with little to no human interaction." "There isn't enough human oversight."	The Automated Scoring Engine (ASE) goes through a rigorous programming process that is led and checked by humans.* For each item being scored: • The ASE uses a sample of ~3,000 human-scored responses from the field test for programming. • TEA evaluates the performance for each item and compares it to human scoring. • TEA evaluates the performance for each item and compares it to how humans would score. • The ASE is monitored throughout the scoring cycle to ensure that it remains calibrated to the anchor set. ◦ Similar to human scorers who need to be constantly calibrated throughout the scoring window, there is a parallel process for the ASE.

Condition Codes

"The Automated Scoring Engine (ASE) is flawed and cannot accurately assess all types of student writing." "The Automated Scoring Engine (ASE) scoring may prioritize quantity over quality, favoring longer responses or specific writing styles that align more closely with its programmed criteria, rather than accurately assessing the content and effectiveness of student writing."	The ASE assigns condition codes to some responses, which are each routed to two trained human scorers.* • Condition codes indicate that a response uses just a few words, uses mostly duplicated text, is written in another language, consists primarily of text from the passage, uses vocabulary that does not overlap with the vocabulary in the subset of responses used to program the ASE, or uses language patterns that are reflective of off-topic or off-task responses. • This routing ensures that unusual responses are scored fairly and accurately. Any response directed to a human scorer is retained as the official score of record.

ASE & Human Scoring Comparison

"The Automated Scoring Engine (ASE) rates harder than human scorers, which will make it harder for our students to get higher scores on their assessments." "The Automated Scoring Engine (ASE) grading will result in fewer students scoring Approaches or higher on RLA STAAR tests."	TEA conducted a proof-of-concept study with STAAR Spring 2023 operational data before implementing hybrid scoring.* • Spring 2023 constructed response items were scored entirely by humans. The study was conducted after score reports were sent out to districts. • The study "re-scored" constructed response items with the Automated Scoring Engine and compared how closely the engine performed to humans. • Five subgroups were examined (Male, Female, Black, Hispanic/Latino, and White). • The proof-of-concept study was successful and found that the Automated Scoring Engine met the performance criteria to be implemented operationally. • A detailed technical report on the Spring 2023 study can be found on the Assessment Reports and Studies webpage[1].

MISCONCEPTIONS	FACTS
ASE Accuracy	
"The Automated Scoring Engine technology does not score student responses as accurately as humans." "AI computers are the sole determinants of ECR scores."	When spring 2023 student responses (which were only scored by humans) were put through the ASE, the human-to-ASE exact score match rate was statistically equal to or greater than the human-to-human exact score match rate.* • The ASE exhibited comparable reliability in grading compared to human scorers assessing one another's work. • Adjacent agreement and score distribution analysis also show that the ASE was functioning as expected. ◦ 2023 STAAR Hybrid Scoring Study[2]
Constructed Response (CR) Zeroes	
"The initial results in some districts across Texas produced more zeroes (0) on constructed response questions."	These were scored appropriately. TEA allowed educators to view the results to see why the CR were scored zeros. Students wrote things that were off-topic or they did not answer the question.^
Rescoring Process	

"We have no recourse if we disagree with the score a response receives from the Automated Scoring Engine (ASE)."
 	
"The Automated Scoring Engine (ASE) scores are final and cannot be disputed."
	Educators and students can request rescoring if they believe there's an error in the initial scoring. Human scorers review these requests and may adjust scores accordingly. The process for rescoring follows these steps:
	• District Testing Coordinator completes a Rescore Request in TIDE (Administering Tests>Appeals/Score Codes).
	• A purchase order number is required when the request is made.
	   ◦ If the student's constructed response score goes up, there is no charge.
	   ◦ If there is no change, the LEA will be charged $50 (a student's score will not go down).
	• If the request is made during the rescore window on the Calendar of Events and the score goes up, any resulting change to the student's overall score/performance level will be reflected in the LEA's accountability calculations.
	• If the request is made after the rescore window and the score goes up, the change will not be reflected in the LEA's accountability calculations but will be reflected in the student's final overall test score (e.g., this could add enough points to the student's score to make it a passing score for the English I EOC).

MISCONCEPTIONS	FACTS
Field Test Scoring	
"There is no human aspect to constructed response scoring." "There isn't enough human oversight." "Everything is based on computers with no human input." "Field tests are administered to develop good items, but only computers are looking at the field test responses."	After the stand-alone field test, each CR item is scored against a rubric by two humans.* - To qualify as a rater, one must have a four-year college degree and experience teaching at the assigned grade level. - Raters undergo rigorous training to learn how to use the standardized rubric to score student responses. - To pass training, they must accurately rate example responses that have already been scored. - Each certified rater's performance is calibrated at regular intervals to ensure that all responses are graded consistently across Texas.
Communication Timeline	
"TEA hasn't communicated this to us." "There was no advance notice of this change."	TEA communicated the transition with the public starting Fall of 2023.* ESC Region 13 communicated this transition throughout the 2023-2024 school year: - DTC training in October 2023 and March 2024 - STAAR Studded events in January 2024 and February 2024 - Content specialist training throughout the year - Email communication

Sources:

*https://tea.texas.gov/student-assessment/testing/hybrid-scoring-key-questions.pdf

[1] https://tea.texas.gov/student-assessment/testing/student-assessment-overview/assessment-reports-and-studies

[2] https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2023-staar-hybrid-scoring-study.pdf

^https://www.kvue.com/article/news/education/schools/texas-staar-test-ai-grading-portions/269-d86101c7-bdf2-4328-9726-a9797979c046