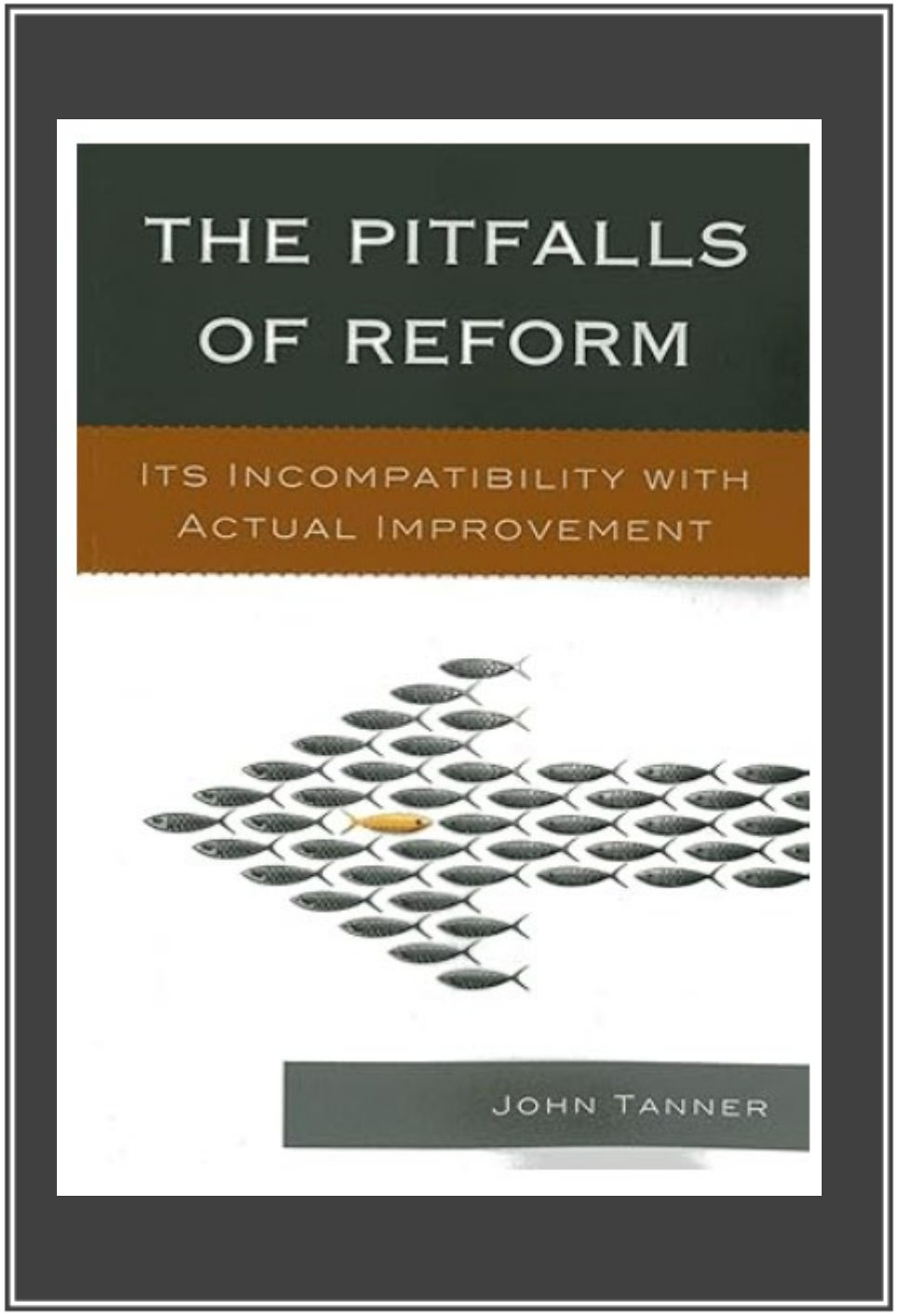
ACCOUNTABILITY & ASSESSMENT UPDATE

Benefits-Based Accountability – featuring John Tanner, bravEd

STAAR Constructed Response Scoring (ASE vs Craig's List)

TEXAS SCHOOL ALLIANCE



FAQ:

How are Texas' A-F accountability letter grades calculated?

"School grading is not clear, simple or transparent… it creates confusion among educators, and fails to offer the public useful or accurate information about their schools. Educators can't explain why a school earned a C or D without referring to a ~~60~~ 200-page technical manual."

John Tanner, The Pitfalls of School Grading,
2016 TASA/TASB presentation

TEXAS OBSERVER

SINCE 1954

TEST SCORE WOES IN LEWISVILLE COULD SIGNAL MORE STAAR TROUBLE ACROSS TEXAS

School district will spend over $50,000 to check for more erroneous scores.

by **PATRICK MICHELS**

APRIL 20, 2016, 2:27 PM, CDT

Weeks after more than 14,000 students across Texas found that their answers on a state-required test had been **accidentally deleted**, administrators in Lewisville schools are working to determine the scope of a separate problem. In this case, some high school English students' short handwritten answers were incorrectly scored zero, jeopardizing their chances to graduate.

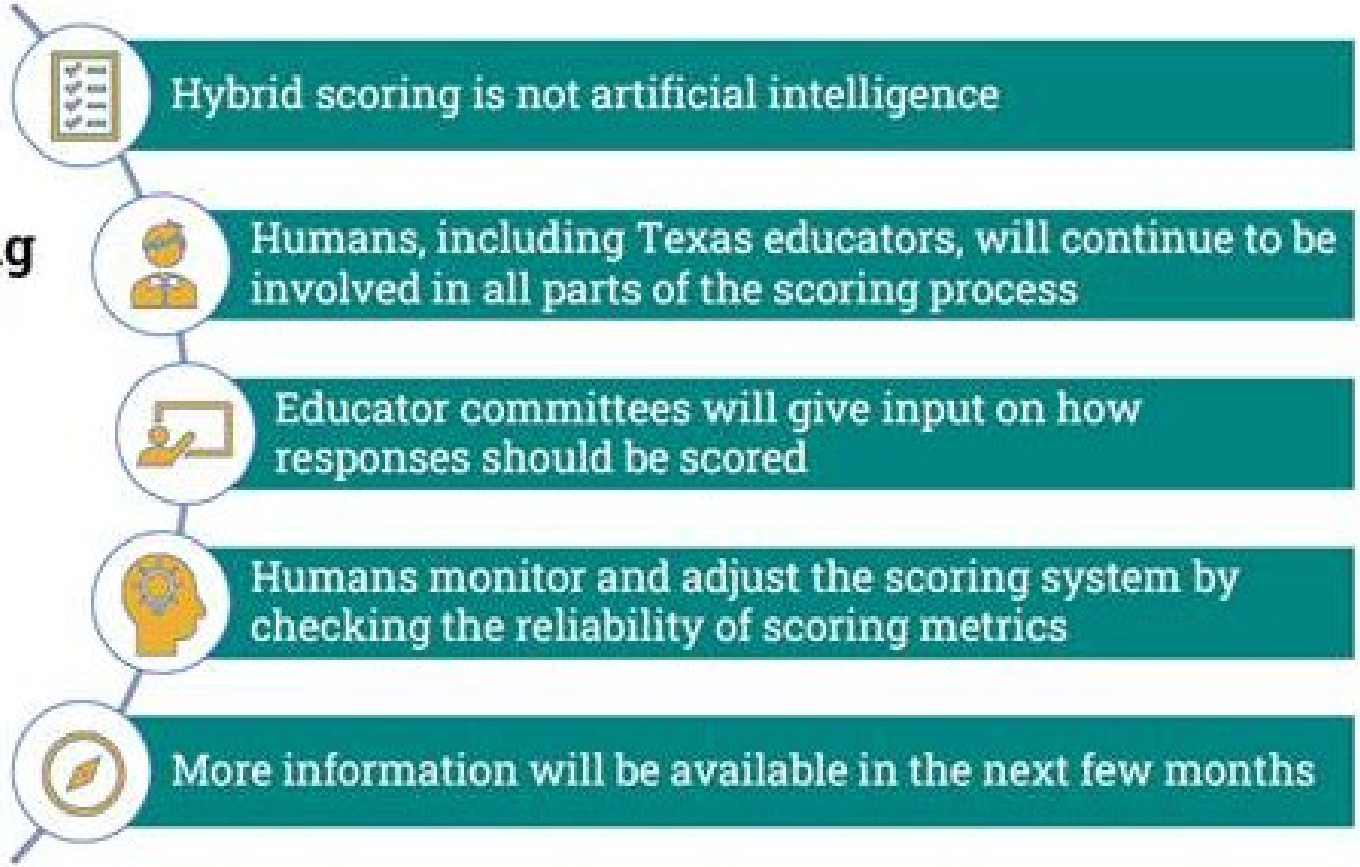
The district first suspected problems when administrators noticed that 17 percent of its English I test-takers received zero scores on their short-answer questions, double the rate of zeros on the previous year's test. When they compared students' scores on the short answer section to their scores on the longer essay questions, they found 76 percent of the students who scored zero on short answer questions also scored passing grades on their essays.

Lewisville has already paid the state's test vendor, Educational Testing Service, $3,225 to review the suspicious scores, 20 percent of which came back with a higher grade. On Monday, the Lewisville Independent School District board approved spending up to $48,250 more to review all of the tests it believes could have been mis-scored.

The *Lewisville Texan Journal* **reported Monday** that board members were, in the words of one member, "flat ticked off" about footing the bill to have ETS double-check its own grading.

TEA began implementing a hybrid-scoring model in December 2023.

TEA began implementing a hybrid-scoring model that incorporates automated scoring alongside our human expert scorers in December 2023.



Source: TEA Assessment Presentation TCTELAR Conference, slide 56. 1-26-2024

Overview of Scoring

Student responses to short constructed-response (SCR) questions and extended constructed-response (ECR) questions are scored using a hybrid scoring model. This means that the responses are initially scored by an automated scoring engine (ASE) and at least 25 percent of student responses are then routed to human scorers. The scores generated by human scorers are used throughout the scoring window to monitor scores that are being generated by the ASE. This process is similar to a second reader in human scoring models.

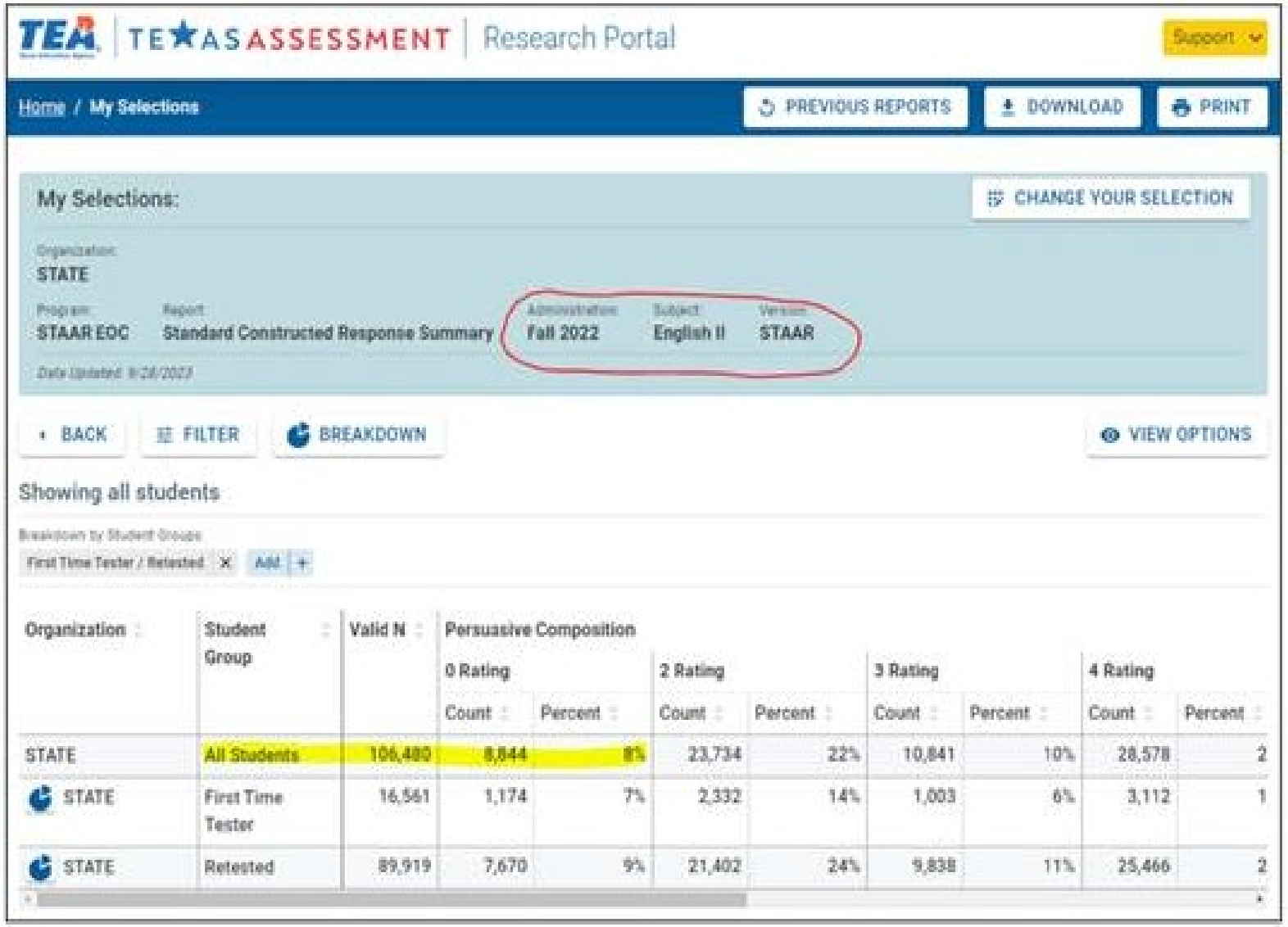
Training the ASE

The ASE is trained on student responses and human scores from the field-test data. It is trained to emulate how humans would score student responses for each constructed-response question. TEA evaluates the performance of the ASE for each question using a subset of student responses and corresponding scores that the ASE has not processed. Performance is examined relative to how humans would score the responses to each question; TEA requires the ASE to agree with

---

Scoring Process for STAAR Constructed Responses

human scorers at the same rate human scorers agree with one another and that the distribution of ASE scores is similar to the distribution of human scores.



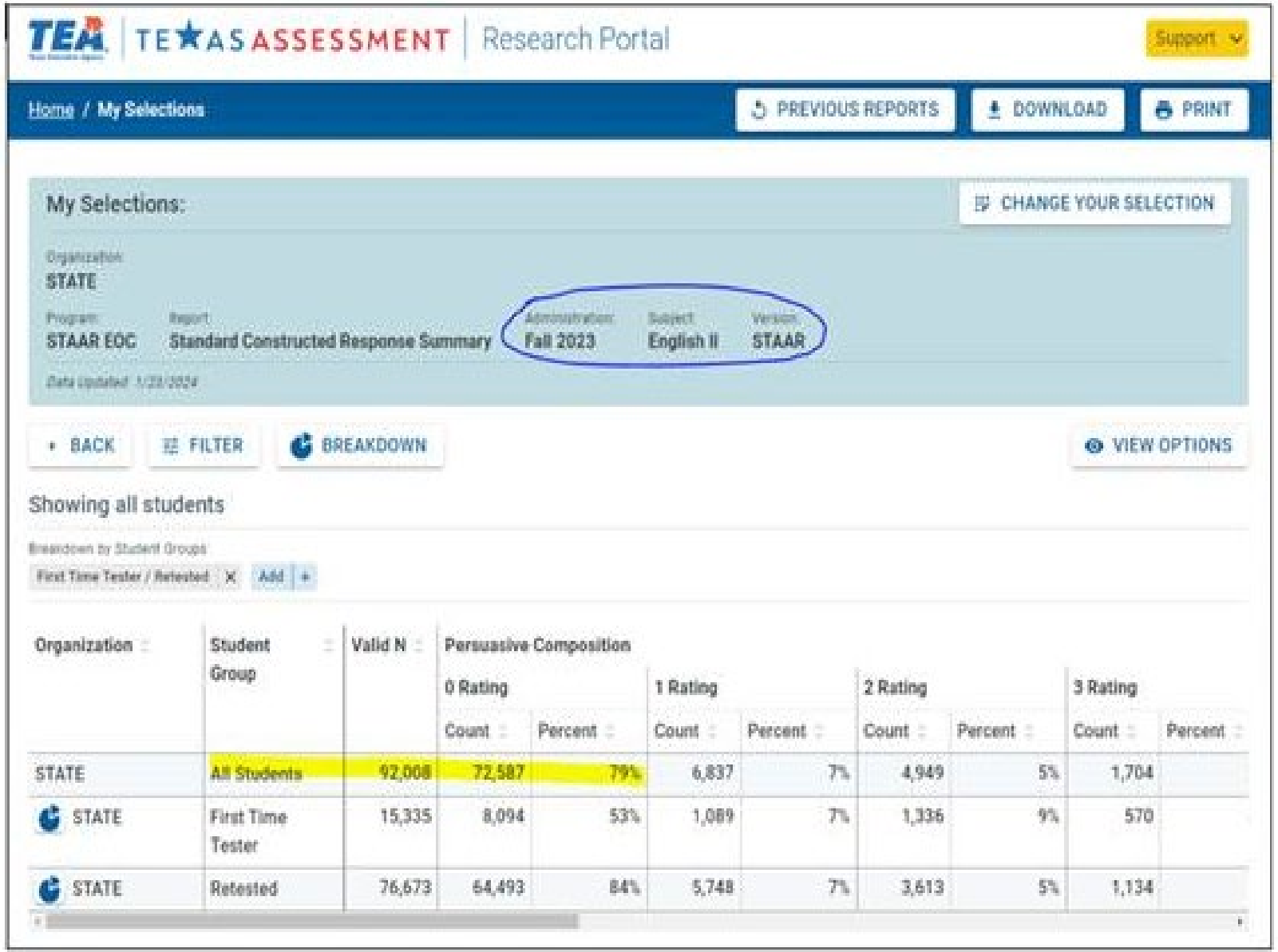
Organization	Student Group	Valid N	Persuasive Composition							
			0 Rating		2 Rating		3 Rating		4 Rating	
			Count	Percent	Count	Percent	Count	Percent	Count	Percent
STATE	All Students	106,480	8,844	8%	23,734	22%	10,841	10%	28,578	2
STATE	First Time Tester	16,561	1,174	7%	2,332	14%	1,003	6%	3,112	1
STATE	Retested	89,919	7,670	9%	21,402	24%	9,838	11%	25,466	2

DEC 2022 English II EOC

TEA used Human Scorers

8% of students scored 0 out of 8 pts.

https://txresearchportal.com/



Organization	Student Group	Valid N	Persuasive Composition							
			0 Rating		1 Rating		2 Rating		3 Rating	
			Count	Percent	Count	Percent	Count	Percent	Count	Percent
STATE	All Students	92,008	72,587	79%	6,837	7%	4,949	5%	1,704	
STATE	First Time Tester	15,335	8,094	53%	1,089	7%	1,336	9%	570	
STATE	Retested	76,673	64,493	84%	5,748	7%	3,613	5%	1,134	

DEC 2023 English II EOC

TEA used AI / Hybrid Scoring Model / ASE

79% of students scored 0 out of 10 pts.

https://txresearchportal.com/

TEA | TEXAS ASSESSMENT | Research Portal

Support

Home / My Selections

PREVIOUS REPORTS | DOWNLOAD | PRINT

My Selections:

CHANGE YOUR SELECTION

Organization:
[redacted] ISD

Program: STAAR EOC
Report: Standard Constructed Response Summary
Administration: Fall 2022
Subject: English II
Version: STAAR

Data Updated: 9/28/2023

BACK | FILTER | BREAKDOWN | VIEW OPTIONS

Showing all students

Breakdown by Student Groups:
First Time Tester / Retested ✕ | Add +

Organization	Student Group	Valid N	Persuasive Composition							
			0 Rating		2 Rating		3 Rating		4 Rating	
			Count	Percent	Count	Percent	Count	Percent	Count	Percent
[redacted] ISD	All Students	643	45	7%	184	29%	72	11%	155	2
[redacted] ISD	First Time Tester	57	7	12%	10	18%	4	7%	10	1
[redacted] ISD	Retested	586	38	6%	174	30%	68	12%	145	2

TEA Texas Education Agency | TEXAS ASSESSMENT | Research Portal

Support

Home / My Selections

PREVIOUS REPORTS | DOWNLOAD | PRINT

My Selections:

CHANGE YOUR SELECTION

Organization: [redacted] SD

Program: STAAR EOC

Report: Standard Constructed Response Summary

Administration: Fall 2023

Subject: English II

Version: STAAR

Data Updated: 1/23/2024

BACK | FILTER | BREAKDOWN | VIEW OPTIONS

Showing all students

Breakdown by Student Groups:

First Time Tester / Retested × | Add +

Organization	Student Group	Valid N	Persuasive Composition							
			0 Rating		1 Rating		2 Rating		3 Rating	
			Count	Percent	Count	Percent	Count	Percent	Count	Percent
[redacted] SD	All Students	695	598	86%	40	6%	27	4%	9	
[redacted]	First Time Tester	81	60	74%	8	10%	3	4%	0	
[redacted]	Retested	614	538	88%	32	5%	24	4%	9	

Organization	Administration	STAAR					
		Tests Taken	Average Scale Score	Performance Levels			
				Did Not Meet	Approaches and Above	Meets and Above	Masters
[redacted] ISD	Fall 2022	643	3531	78% 499 Tests	22% 144 Tests	7% 46 Tests	0% 1 Test

Organization	Administration	STAAR					
		Tests Taken	Average Scale Score	Performance Levels			
				Did Not Meet	Approaches and Above	Meets and Above	Masters
[redacted] ISD	Fall 2023	695	3586	76% 528 Tests	24% 167 Tests	9% 62 Tests	0% 1 Test

Scoring Process for STAAR Constructed Responses

Constructed-Response Questions

The State of Texas Assessments of Academic Readiness (STAAR®) is a standardized academic achievement test designed to measure the extent to which a student has learned and is able to apply the defined knowledge and skills of the Texas Essential Knowledge and Skills (TEKS) for each tested grade, subject, and course. STAAR includes open-ended questions that reflect classroom test questions and allow students more ways to show their understanding of the TEKS. Two specific types of open-ended questions, known as extended constructed-response (ECR) questions and short constructed-response (SCR) questions, require students to draft written responses. ECR questions are included on all STAAR reading language arts (RLA) assessments—grades 3–8, English I, and English II. SCR questions are included on:

- grades 3–8 RLA,
- grades 5 and 8 science,
- grade 8 social studies,
- English I,
- English II,
- Biology, and
- U.S. History.

Student responses to SCR and ECR questions are evaluated using a holistic scoring process. The responses are evaluated as a whole based on pre-established criteria identified and detailed in the scoring rubrics. The scoring rubrics are used to determine the effectiveness of each written response. This document outlines the scoring process the Texas Education Agency (TEA) uses for STAAR SCRs and ECRs.

What questions (or answers) do you have?

Scoring Process for STAAR Constructed Responses

Constructed-Response Questions

The State of Texas Assessments of Academic Readiness (STAAR®) is a standardized academic achievement test designed to measure the extent to which a student has learned and is able to apply the defined knowledge and skills of the Texas Essential Knowledge and Skills (TEKS) for each tested grade, subject, and course. STAAR includes open-ended questions that reflect classroom test questions and allow students more ways to show their understanding of the TEKS. Two specific types of open-ended questions, known as extended constructed-response (ECR) questions and short constructed-response (SCR) questions, require students to draft written responses. ECR questions are included on all STAAR reading language arts (RLA) assessments—grades 3–8, English I, and English II. SCR questions are included on:

- grades 3–8 RLA,
- grades 5 and 8 science,
- grade 8 social studies,
- English I,
- English II,
- Biology, and
- U.S. History.

Student responses to SCR and ECR questions are evaluated using a holistic scoring process. The responses are evaluated as a whole based on pre-established criteria identified and detailed in the scoring rubrics. The scoring rubrics are used to determine the effectiveness of each written response. This document outlines the scoring process the Texas Education Agency (TEA) uses for STAAR SCRs and ECRs.

Overview of Scoring

Student responses to short constructed-response (SCR) questions and extended constructed-response (ECR) questions are scored using a hybrid scoring model. This means that the responses are initially scored by an automated scoring engine (ASE) and at least 25 percent of student responses are then routed to human scorers. The scores generated by human scorers are used throughout the scoring window to monitor scores that are being generated by the ASE. This process is similar to a second reader in human scoring models.

NOTE: Student responses to SCR and ECR questions for STAAR Spanish assessments are 100 percent human scored and do not get scored by the ASE.

Pre-scoring Activities

Before scoring begins, a few activities occur which optimize how the Texas Education Agency (TEA) assesses the TEKS on STAAR.

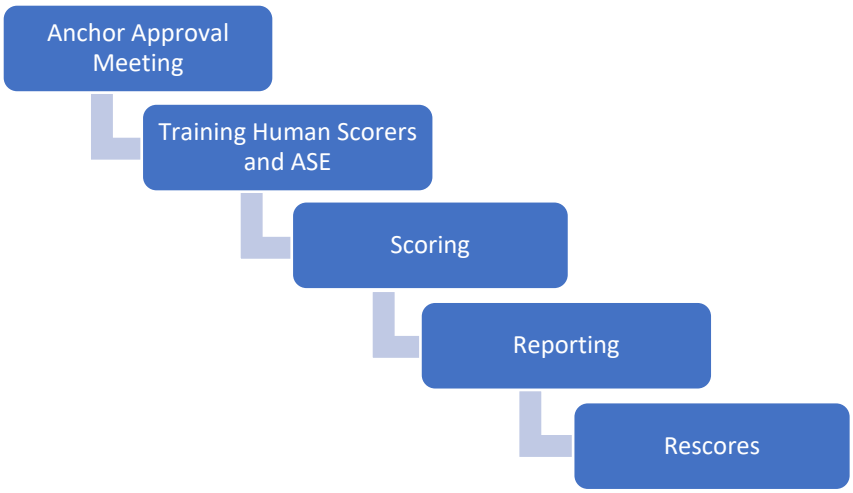
- Developing New Questions
  Development activities begin with content and assessment experts developing the passages and questions that will assess the TEKS. Committees of Texas educators (e.g., teachers, instructional specialists, district coordinators, principals, etc.) review passages and questions and provide input

and feedback. Once passages and questions are reviewed and edited accordingly, they are approved by TEA and are ready for field testing.

- Field Testing New Questions
  Field testing questions is the process of gathering performance data on test questions, but it does not impact students' scores on STAAR. TEA conducts field testing of newly developed questions each year either through embedded field-test questions or through stand-alone field tests.

Scoring Activities

The steps outlined below each have an important role in the overall scoring process.



Step 1: Anchor Approval Meeting

A selected group of student responses are collected during field testing and are scored using scoring rubrics, i.e., the TEA-approved scoring criteria. Student responses from the field-tested questions are taken to anchor approval meetings. The purpose of anchor approval meetings is to apply the scoring rubrics to sample student responses and establish scoring boundaries. Student responses used to establish scoring boundaries are called "anchor responses", and these responses clearly exemplify each score point in the rubric. During the anchor approval meetings, educators identify anchor responses, provide consensus scores for these responses, and discuss the rationale for the scores these responses receive. These discussions are used to develop training notes or annotations for the responses. After the meeting, item sets intended for use as practice and qualification sets are created based on the approved anchor responses. Educators are trained and qualify as scorers using these item sets.

Step 2: Training Human Scorers and ASE

Both human scorers and the ASE are trained with TEA-approved student responses.

- Training Human Scorers
  Scorers are assigned to score specific grade/subject or course SCRs or ECRs based on their experience and educational background. For each question type, human scorers go through a

similar process. They study and review the anchor responses. Then scorers are given the opportunity to practice on the different ways students are responding to the question. Scoring directors and scoring supervisors are available for clarification and guidance during the process and support the scorers in their training as needed.

After completing practice sets, human scorers must demonstrate a complete understanding of the rubrics by successfully completing a qualification set. The qualification set is a quality measure to ensure that scorers are following the rubric and can accurately apply scores. Only scorers who complete the training and pass the qualification set may score student responses. Scorers who do not pass the qualification set are not allowed to score student responses for STAAR.

- RLA SCR Training
  The SRCs for the writing section of the RLA assessment are scored using a generic 0- or 1-point rubric. The SRCs for the reading section use an item-specific 0- to 2-point rubric. Human scorers are trained to apply the 1-point rubric or 2-point rubric using anchor responses or anchor sets that include student responses at each score point. Scorers then complete practice sets where they score student responses for their assigned grade or course. Both the anchor and practice sets include annotations explaining why and how the score was assigned. The scoring rubric used for each SCR question on a STAAR released test can be found in the constructed response scoring guides available on the STAAR Reading Language Arts Resources webpage.
- RLA ECR Training
  The ECRs for RLA are scored using an item-specific 5-point rubric that identifies scores based on two traits—development and organization of ideas (up to 3 points) and language conventions (up to 2 points). Human scorers are trained to apply the 5-point rubric using an anchor set that includes student responses at each score point. Scorers then complete practice sets where they score student responses for their assigned grade or course. Both the anchor and practice sets include annotations explaining why and how the score was assigned. The STAAR RLA scoring rubrics for ECRs are available on the STAAR Reading Language Arts Resources webpage.
- Science and Social Studies SCR Training
  Both science and social studies SCRs are scored using 0- to 2-point item-specific rubrics. Human scorers are trained to apply the rubric by using the approved anchor sets and then must score at least two practice sets. Both the anchor and practice sets include annotations explaining why and how the score was assigned. The scoring rubric used for each SCR question on a STAAR released test can be found in the constructed response scoring guides available on the corresponding STAAR Science Resources and STAAR Social Studies Resources webpages.

- Training the ASE
  The ASE is trained on student responses and human scores from the field-test data. It is trained to emulate how humans would score student responses for each constructed-response question. TEA evaluates the performance of the ASE for each question using a subset of student responses and corresponding scores that the ASE has not processed. Performance is examined relative to how humans would score the responses to each question; TEA requires the ASE to agree with

human scorers at the same rate human scorers agree with one another and that the distribution of ASE scores is similar to the distribution of human scores.

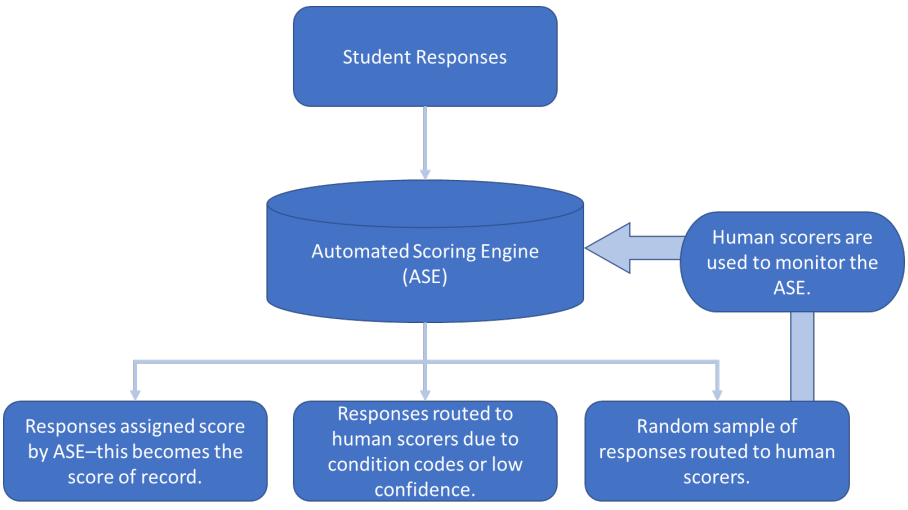
As part of the training process, the ASE calculates confidence values that indicate the degree to which the ASE is confident the score it has assigned matches the score a human would assign. The ASE also identifies student responses that should receive condition codes. Condition codes indicate that a response is blank, uses too few words, uses mostly duplicated text, is written in another language, consists primarily of stimulus material, uses vocabulary that does not overlap with the vocabulary in the subset of responses used to train the ASE, or uses language patterns that are reflective of off-topic or off-task responses.

Step 3: Scoring

All constructed-response questions for the English language version of STAAR are scored using a hybrid scoring model. All constructed-response questions for the Spanish language version of STAAR are scored using only human scorers.

The process for the hybrid scoring model is outlined below.

- All student responses are scored for the first time by the ASE.
- At least 25 percent of student responses for each grade/subject and course are then routed to trained human scorers. This set of second scores by human scorers is used to monitor the ASE scoring and is similar to what is done in human scoring to examine how often scorers agree with one another.
- Student responses that the ASE assigns condition codes (e.g., use words not seen in the responses used to train the ASE) or that are identified as “low confidence” are routed to trained human scorers. The low confidence responses are often those responses that are on the border between two score points. The purpose of this routing is to ensure that unusual or borderline responses receive fair and accurate scores.



Any student responses that are routed for human scoring maintain the score assigned by the human score as the score of record.

Throughout the scoring window, both ASE and human scoring results are monitored daily to ensure that scores consistently align to the scoring rubrics.

Step 4: Reporting

Results from STAAR administrations are sent to districts after all testing is complete. Reports for each student include the total number of points earned, the scale score, and the performance level. Item-level reports include the average number of points earned for each question on the assessment, including constructed-response questions. For more information about reporting of STAAR results, refer to the Interpreting Results page in the *District and Campus Coordinator Resources*.

Step 5: Rescores

If district personnel or a parent or guardian has concerns about a student's score on a constructed-response question, district testing personnel can request that the student's response be rescored for a fee. Rescore requests are sent for human scoring. The results of the rescore are provided to district testing personnel. If the score on the constructed-response question changes, the fee is waived.

The rescore request must be submitted in the Test Information Distribution Engine (TIDE) by district testing personnel within a specific window. Refer to the Calendar of Events for specific dates. For assistance in submitting rescore requests in TIDE, contact Texas Testing Support at (833) 601-8821 or TexasTestingSupport@cambiumassessment.com.