The State of Texas Assessments of Academic Readiness (STAAR®) Hybrid Scoring Study Methods and Results: Spring 2023 Items

Table of Contents

Executive Summary

This study examines the robustness of the models programmed with the field-test data on the operational responses and scores using Spring 2023 State of Texas Assessments of Academic Readiness (STAAR®) items and data. The study found that the tested Automated Scoring Engine met sufficient performance criteria on each of the field-test programmed and operationally programmed held-out validation samples. We expect that the results of the study will generalize to future administrations starting in December 2023.

Models programmed on the sample of the first 50% of the operational data in the program met sufficient performance criteria for all students and for the five student groups evaluated: Male, Female, Black, Hispanic/Latino, and White.

Under current administration conditions, models will need to be reprogrammed on the operational data during the scoring window. Moving forward, we will plan that all models for all items be reprogrammed on the operational data, regardless of performance. Once deployed, all responses are rescored with the newly reprogrammed model and with any newly determined condition codes or low confidence responses routed for human scoring.

An automated score will never be used in the human scoring process; rather, any response routed for human scoring or hand-scoring use spring 2023 hand-scoring rules to assign scores. This approach will better support engine reprogramming on the operational data allowing for a direct comparison of the engine to humans. Additionally, any scores assigned during human scoring are considered the score of record; all other responses receive the score assigned by the reprogrammed model.

Additional work will continue to further refine the condition codes and thresholds used. These will likely vary by item type and grade and will undergo review and analysis to ensure alignment to the scoring rubric and human-assigned condition codes.

Introduction

The purpose of this study is to examine the robustness of automated scoring models on the STAAR operational data and to conduct further analysis on the modeling to prepare for the operational hybrid automated/human scoring starting in December 2023 for STAAR end-of-course (EOC) assessment items and continuing through spring 2023 for STAAR grade 3–8 items. Spring 2023 items and data were used to examine engine performance in an operational setting under the assumption that these results will generalize to future administrations. The benefit of this approach is that all responses received human scores in spring 2023, so various configurations around hybrid automated/human scoring can be examined. It is important to note that STAAR items administered as part of the spring summative assessment are released that same year, and so are not available for future use. This means that the models programmed in this study cannot be used in any future administration.

The key elements of the approach are as follows. First, we assume that the initially deployed models are programmed on field-test data. These models are then evaluated for performance on a random verification sample that has been routed for human scoring. If models perform well against key criteria, we can assume they are performing adequately across all responses. Otherwise, models need to be reprogrammed on the operational data and then redeployed to score all new and any previously scored responses. The analysis approach involves multiple phases.

- In Phase 1, models are programmed on the field-test data for all spring 2023 items and performance is evaluated using key criteria.
- In Phase 2, models are applied to a representative sample of spring 2023 operational data, and routing categories that mimic potential operational use are defined. These categories are automated scoring engine-assigned condition codes, random verification sample (15%), responses with low confidence values (10%), and all remaining responses. It is assumed that the first three categories are either routed for human review or are accurately scored. The last category is assumed to be rubric-valid responses for which the automated scoring engine would be the sole scorer.
- In Phase 3, the performance of the engine is examined on the random verification sample and evaluated relative to the key criteria.
- In Phase 4, models are reprogrammed on the first 50% of the verification sample and evaluated relative to key criteria.
- In Phase 5, reprogrammed models are applied to the remaining responses, and any new low confidence and automated scoring engine-assigned condition code responses are identified.
- In Phase 6, the performance of the reprogrammed engine is examined on the remaining responses.

The results of the study will be used to make recommendations around the adequacy of autoscoring in the STAAR program and to inform decisions around the hybrid automated/human scoring configurations. They also illustrate the proportion of items needing to be reprogrammed within the operational scoring window.

Method

Items

The items administered in each program are presented in Table 1 along with information on the program, item type, rubric score points, and number of responses in the field test and operational test (OP). There are 27 STAAR items. Eight items are extended constructed-response (ECR) items, and the remaining 19 items are short constructed-response (SCR) items. The number of responses available for programming the automated scoring engine from the STAAR Stand-Alone Field Test (SAFT) in 2022 is quite low for most reading language arts (RLA) items (ranging from 1,182 to 2,871), particularly ECRs, compared to what is recommended (2,500–4,000).[1] The science and social studies items have sample sizes recommended for engine programming with the exception of items 71344 and 60001.

Table 1. STAAR Open-Ended Items Administered in Spring 2023

Grade	Item ID	Type	Program	Subject	Score points	SAFT N	OP N
3	79024	SCR	STAAR 3–8	RLA	1	2431	354824
4	78742	SCR	STAAR 3–8	RLA	1	2570	363053
5	80822	SCR	STAAR 3–8	RLA	1	1182	371728
6	80104	SCR	STAAR 3–8	RLA	1	2578	389822
7	81164	SCR	STAAR 3–8	RLA	1	1240	399038
8	79244	SCR	STAAR 3–8	RLA	1	1334	408725
9	80399	SCR	STAAR EOC	RLA	1	2871	511507
9	81260	SCR	STAAR EOC	RLA	1	1279	465307
8	73863	SCR	STAAR 3–8	RLA	2	1289	408888
9	68311	SCR	STAAR EOC	RLA	2	1436	511938
9	70928	SCR	STAAR EOC	Biology	2	4339	455242
9	70937	SCR	STAAR EOC	Biology	2	2542	459605
5	71344	SCR	STAAR 3–8	Science	2	1656	388092
8	60001	SCR	STAAR 3–8	Science	2	1978	405398
8	74531	SCR	STAAR 3–8	Science	2	2742	406040
9	55826	SCR	STAAR EOC	U.S. History	2	2474	375359
9	72841	SCR	STAAR EOC	U.S. History	2	6510	375457
8	72436	SCR	STAAR 3–8	Social Studies	2	6099	411882
8	72439	SCR	STAAR 3–8	Social Studies	2	2493	411544
3	55391	ECR	STAAR 3–8	RLA	Conv: 2, Ideas: 3	1259	355087
4	12632	ECR	STAAR 3–8	RLA	Conv: 2, Ideas: 3	1289	363427
5	12638	ECR	STAAR 3–8	RLA	Conv: 2, Ideas: 3	1165	371737
6	12666	ECR	STAAR 3–8	RLA	Conv: 2, Ideas: 3	1291	389680

[1] This range is recommended in order to obtain responses for rarer score points and to ensure a sample size of 375 or greater for validating the engine.

Grade	Item ID	Type	Program	Subject	Score points	SAFT N	OP N
7	73118	ECR	STAAR 3–8	RLA	Conv: 2, Ideas: 3	1420	398709
8	73991	ECR	STAAR 3–8	RLA	Conv: 2, Ideas: 3	1353	407627
9	68583	ECR	STAAR EOC	RLA	Conv: 2, Ideas: 3	1400	509294
9	68776	ECR	STAAR EOC	RLA	Conv: 2, Ideas: 3	1267	463011

Note. ECR scores are reported to educators on a summed scale using the resolution rules outlined for hand-scoring.

Hand-scoring

Hand-scoring rules varied by administration and item type. Table 2 and Table 3 outline the scoring rules associated with each administration and item type.

All field-test scoring was conducted in spring and summer 2022. For the field test, all items underwent scoring by two independent readers with resolution of non-exact scores. The purpose of this approach was to obtain high-quality scores with which to program the engine and to calibrate item parameters.

All spring 2023 operational scoring occurred within the defined operational scoring windows. For the spring 2023 operational tests, SCRs had only a portion of responses (25%) routed for a second read; any non-exact scores for this subset were routed for expert resolution. ECRs had two independent readers with resolution of non-adjacent scores.

Table 2. 2022 Field-Test Hand-Scoring Specifications

Program	Item Type	Scoring Specification
STAAR	RLA 1-point SCR	1. All SAFT Writing FT SCRs will be scored by at least 2 scorers. 2. Resolution is required for any scores not in perfect agreement.
STAAR	RLA, Social Studies, Science 2-point SCR	1. All SAFT Reading, Social Studies, and Science SCRs will be scored by at least 2 scorers. 2. Resolution is required for any scores not in perfect agreement.
STAAR	ECRs	1. All Writing FT ECRs will be scored by at least 2 scorers. Score points by trait: 0–3 for ideas and 0–2 for conventions. 2. Resolution is required for any non-adjacent scores by trait.

Table 3. Spring 2023 Operational Hand-Scoring Specifications

Program	Item Type	Scoring Specification
STAAR	RLA 1-point SCR	1. Writing SCRs will be scored by one scorer (R1). 2. 25% will get a second score (R2). 3. Resolution (R3) is required for any scores not in perfect agreement. R3 is the final score if present. 4. R1/R2 or R3 score of record is the final score and should not be doubled.
STAAR	RLA, Social Studies, Science 2-point SCR	1. Reading, Social Studies, and Science SCRs will be scored by one scorer (R1). 2. 25% will get a second score (R2). 3. Resolution (R3) is required for any scores not in perfect agreement. R3 is the final score if present. 4. R1/R2 or R3 score of record is the final score and should not be doubled.

Program	Item Type	Scoring Specification
STAAR	ECRs	1. 100% of responses will be independently scored by 2 readers: Reader 1 (R1, Initial Read) and Reader 2 (R2, Reliability Read). 2. Scorer agreement will be evaluated **at the trait level** to determine whether a third resolution read (R3, Resolution Read) is required. a. If R1 and R2 have exact agreement or adjacent numerical scores, no further reads will be conducted. b. If R1 and R2 scores are not adjacent, or one reader assigns a numerical score and the other a condition code, the response is sent for R3. c. If R1 and R2 assign different condition codes, the response is sent for R3. d. All responses that are identified as nonscorable and assigned a condition code (excluding blanks) are routed for verification by a supervisor or scoring director. 3. If R3 is not present, then a final trait score is calculated by adding R1 score and R2 score. If R3 score is present, the R3 score is doubled for the final score.

The human assigned condition codes for STAAR appear in Table 4.

Table 4. STAAR Human-Assigned Condition Codes

B = Blank T = Off Topic I = Indecipherable U = Illegible F = Written in a language other than tested language D = Insufficient response C = Lacks any original writing P = Does not write in prose R = Refuses to write

Description of the Automated Scoring Engine

The automated scoring engine has demonstrated capability to produce comparable results to humans across a variety of item types and grades. The use of automated scoring helps to ensure that scores are returned quickly, at lower cost, and ensures consistency in scoring within and across administrations. This technology is currently used in six state-level summative programs and fourteen state-level interim programs.

The automated scoring engine uses features associated with writing quality and features associated with response meaning. Writing quality features include measures of syntax, grammatical/mechanical correctness, spelling correctness, text complexity, paragraphing quality, and sentence variation and quality. We build two models in parallel and combine the outputs of these models to predict the response score. Ensembling generally produces better performance than the use of a single model (Zhou et al., 2002).

The automated scoring engine also produces condition codes and confidence values as part of its scoring process. Each method is useful in identifying non-attempts, unusual responses, or

borderline responses that can be routed for human verification scoring. Condition codes include those indicating that a response is blank, uses too few words, uses mostly duplicated text, is written in another language, consists primarily of stimulus material or the stem, uses vocabulary that does not overlap with the sample used to program the engine, or uses language patterns that are reflective of the set of human-assigned condition code responses (Table 5). The automated scoring engine also has a condition code that can be used to capture unusual scoring patterns, such as very short responses receiving scores greater than the lowest rubric score. Finally, the thresholds for the engine condition codes are undergoing continual review and will be revised when necessary.

Table 5. STAAR Automated Scoring Engine Condition Codes

Condition Code	Description	Threshold
No Response	Response was empty or consisted only of white space (space characters, tab characters, return characters).	N.A.
Duplicate Text	Response contains a significant amount of duplicate or repeated text.	0.7
Prompt Copy Match	Response consists primarily of text from the passage.	0.9 SCRs, 0.8 ECRs
Non-Scorable Language	Response is in Spanish (must be at least 30 characters).	N.A.
Refusal to Respond	Response is of the form “I don’t know” or “I don’t care.”	N.A.
Not Enough Data	Response has too few words to be considered a valid attempt.	Varies by SCRs, 11 for ECRs
Unusual Vocabulary	Most words in the response do not appear in typical responses.	0.5
Non-Specific	Response displays characteristics of condition codes assigned by humans that do not fall under the above condition code categories.	N.A.
Unusual Score	Response has unusual score pattern characteristics (i.e., non-adjacent scores in related dimensions, high scores for brief responses).	ECRs only

The confidence value reflects the degree to which the automated scoring engine is confident in the score it has predicted. A high confidence value indicates that the engine is confident in its prediction; a low confidence value indicates that the engine is less confident in its prediction. The confidence values are calculated as percentiles. The confidence model is programmed (using probit regression) to predict whether the engine score matches the final human score on the held-out validation sample (1=match, 0=non-match) using the patterns of model outputs as predictors. A model is programmed for each dimension; if there are multiple dimensions as with ECRs, the confidence outputs are standardized and summed to provide an overall item confidence score.

Model Programming

CAI programs models for each item and dimension. Data are divided into programming and held-out validation sets with 85% of responses used to program the engine and 15% used to evaluate the engine performance. Data are stratified on the final, resolved score to ensure that score point distributions are evenly represented in both sets. Human-assigned condition codes are removed

prior to programming the models and are added later in the process when applying the automated scoring engine condition codes.

For SCRs, models are programmed on the final, resolved score arising out of the hand-scoring process. For ECRS because of the summed score approach used in hand-scoring, two different models are programmed independently—one on human rater 1 and one on human rater 2. These scores are combined within dimension to produce a summed score. This summation occurred on the probabilities whereby the probability of the summed score is the sum of the products of the model probabilities for all possible sums of the summed score. For example, the probability of a summed score of 2 is the sum of the following products: $P_{model1}(0)*P_{model2}(2)+P_{model1}(1)*P_{model2}(1)+P_{model1}(2)*P_{model2}(0)$. The final score in the summed scale is the argmax of the probabilities, or the score associated with the highest probability.

Evaluation Metrics

Metrics used to examine engine performance are those commonly used in the assessment industry (Williamson, Xi, and Breyer, 2012). These include measures of agreement (Exact Agreement, Quadratic Weighed Kappa [QWK] using Fleiss-Cohen weights) and a distributional measure (Standardized Mean Difference [SMD] using pooled standard deviation).

Definitions for these terms are:

1. *Exact Agreement*—Represents the percentage of responses for which two raters agree on the score. A score of 100% indicates perfect agreement across all responses, and a value of 0% indicates that there was no agreement at all. We expect human-machine (HS-AS) Exact Agreement to be no less than 5.25% of the human-human (H1-H2) exact agreement rate.

2. *QWK*—Also referred to as Cohen's kappa or a kappa value, provides a measure of agreement where a value of 1 represents perfect agreement and a value of 0 indicates random chance. Additionally, QWK considers the magnitude of disagreements. If two raters disagree by more than one score point, (for example, one rater assigns a score of 1 and the other assigns a score of 3) this is penalized more than a disagreement of just one score point. Hence, the term quadratic weighted describes the penalty assigned for such extreme disagreements. We expect item traits will be such that the HS-AS QWK is no less than 0.10 the H1-H2 QWK.

3. *SMD*—Helps us determine if the two rater groups are scoring differently from one another without having to know the scale of a particular test item. To calculate the SMD, we first compute the mean score assigned by each rater. Then we take the difference between the two. In order to obtain a value that can be interpreted across all items, we divide the difference of means by how much variation in scores we see in the entire dataset. A value of 0 indicates that there is no discernible difference in scores assigned by human raters and by an automated scoring model. We expect item traits to differ by no more than a magnitude of 0.15.

CAI used the following thresholds to identify poorly performing items:

- Engine-Final, resolved score exact agreement lower than 5.25% of H1-H2 exact agreement (PARCC, 2015),
- Engine-Final, resolved QWK lower than 0.1 of H1-H2 QWK (Williamson et al., 2012), and
- Engine-Final, resolved SMD magnitude greater than 0.15 (Williamson et al., 2012).

For ECR summed scores, there is no comparable H1-H2 agreement, so only two measures are used:

- Engine-Final, resolved QWK greater than 0.7 (Williamson et al., 2012) and
- Engine-Final, resolved SMD magnitude greater than 0.15 (Williamson et al., 2012).

For each item type and scoring, two sets of results are presented: 1) the average QWK, exact agreement, and SMD values for the two human raters and for the engine and final resolved score; and 2) the percentage of items passing performance thresholds.

Finally, the application of the metrics was conducted on the sample of response in which both automated scoring engine and human-assigned condition codes were removed. This approach was taken because the final determination of the condition codes and thresholds was not yet determined and because the core focus is on the ability of the engine to reproduce human scores.

Field-Test and Operational Scores

Prior to engine programming, the final score distribution data was computed and compared between the field-test and operational hand-scored data. The purpose of this analysis is to identify similarities and differences between the two samples as a first step toward understanding and explaining model performance on the two samples. Models whose programming data are similar in terms of means, standard deviations, and score point distributions relative to the operational data are likely to perform well assuming sufficient programming sample sizes. Models whose programming data differ may result in poorer performance on the operational data because this suggests possible differences in hand-scoring or in how students are responding to items. Assuming the hand-scoring rules and the response characteristics are similar between the field-test and the operational test, the models are likely to perform well.

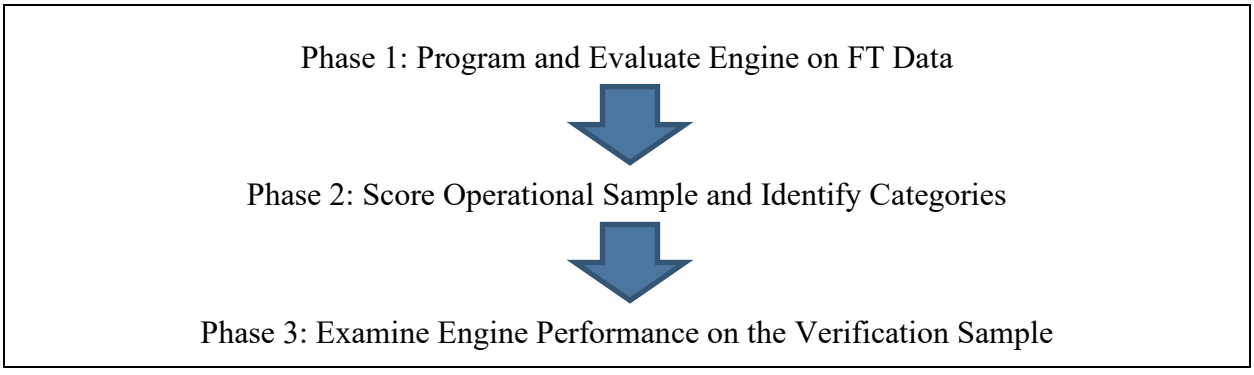
For detailed results, refer to Appendix A: Field-Test and Operational Score Distributions. In STAAR, the operational RLA SCRs generally showed higher mean scores with SMD values ranging from 0.20 to 0.57. Science SCRs showed more varied performance with an SMD range of -0.17 to 0.21 and three items having lower operational means than field-test means. Social studies SCRs also showed varied performance with an SMD range of -0.27 to 0.40 and two items having lower operational means than field-test means. STAAR ECRs showed mostly higher operational means across items and across the two dimensions; the conventions SMDs ranged from -0.01 to 0.54, and the ideas SMDs ranged from 0.08 to 0.56.

These data suggest models programmed on the field-test data alone may not perform well on operational data except for some STAAR RLA SCRs. As a result, TEA uses operational data to train the engine in order to maximize accuracy of student scores.

High-Level Overview of the Study Method

Figure 1 provides a high-level overview of the study method. The Results section below is organized into these phases, and they are reiterated below for clarity given the complexity of the analysis.

- In Phase 1, automated scoring engine models were programmed and evaluated on the field-test data.
- In Phase 2, these programmed models were used to score a sample of operational responses. For STAAR, a random sample of 25% of responses, stratified by administration date, were used. These percentages were chosen to ensure sufficient sample sizes for reprogramming models on the operational data. These field-test programmed models were then used to score the random sample. From these samples, four routing categories were identified: 1) those receiving condition codes by the automated scoring engine, 2) a verification sample of responses chosen to be 15% stratified by date, 3) responses receiving engine confidence percentile values less than 10, and 4) remaining responses. These categories and thresholds were chosen to mimic a possible operational hybrid automated/human scoring approach. In such an approach, the verification sample and low confidence samples would be routed for human scoring. A subset of condition codes would also be routed for human scoring. The responses in category 4, remaining responses, were assumed to be scored only by the engine.
- In Phase 3, the model performance was evaluated on the verification sample.
- In phase 4, models were reprogrammed on the first 50% of the verification sample with 15% of that sample held out for model evaluation. This sample was chosen to mimic the fact that the engine performance would be monitored early in the window and, if not performing well, would be reprogrammed on a substantial portion of the verification sample that was considered reasonably representative of the set of testers throughout the administration.
- In Phase 5, the reprogrammed models were deployed and used to rescore the category 4 responses with any new automated scoring engine condition codes and low confidence responses routed for human scoring.
- In Phase 6, the performance of the engine on responses was examined. In addition, subgroup analysis was computed on student sex (male, female) and student race/ethnicity (White, Black, Hispanic). This analysis used the same metrics outlined for evaluating the automated scoring engine within each subgroup category.



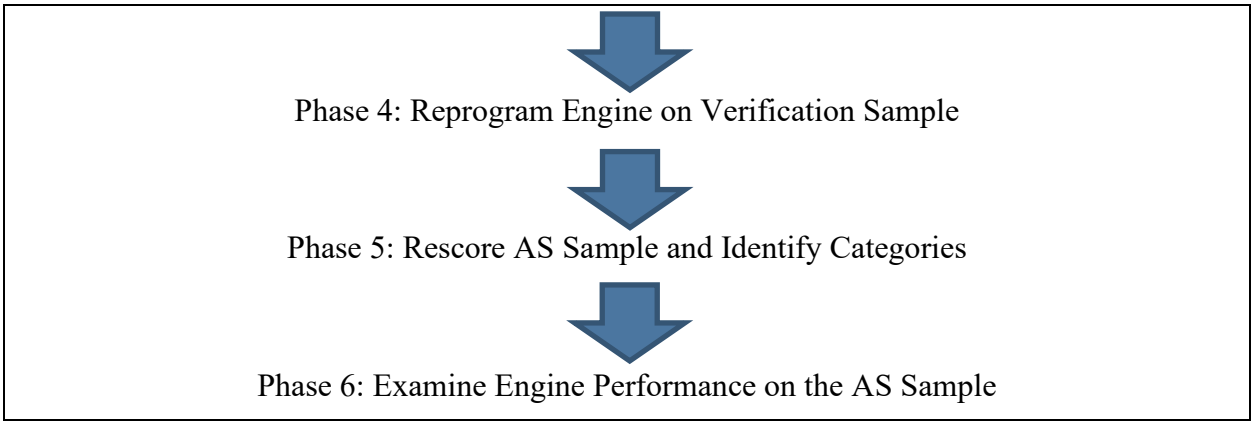


Figure 1. Overview of the Study Method

Results

The results are divided into the six phases of the study.

Phase 1: Field-Test Programmed Model Performance on Field-Test Data

Across the STAAR item types, the automated scoring engine met the performance criteria for almost all items with two items failing at least one criterion (STAAR RLA SCR item 80822 and STAAR ECR item 55391). The results are presented by program and item type. For each program and item type, the QWK, exact agreement, and SMD results are presented first followed by the percentage of items meeting the performance criteria.

Phase 1: STAAR SCRs

The automated scoring engine showed similar QWK and exact agreement with the final, resolved score relative to the two human raters (Table 7). The standardized mean differences between the engine and the final, resolved score ranged from -0.17 (80822) to 0.10 (73863). On average across the items, the engine assigned slightly higher mean scores as indicated by a negative SMD value relative to the final, resolved score. Ninety percent of the RLA SCRs met the three thresholds, and all science and social studies SCRs met the three thresholds (Table 8).

Table 6. Phase 1 AS Performance on STAAR SCRs

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
RLA										
79024	1	330	0.87	0.85	-0.01	93%	93%	-1%	-0.04	0.02
78742	1	359	0.82	0.86	0.04	91%	93%	2%	-0.02	0.03
80822	1	173	0.76	0.77	0.01	88%	90%	1%	-0.02	-0.17

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
80104	1	371	0.59	0.77	0.18	79%	88%	9%	-0.09	-0.03
81164	1	179	0.70	0.79	0.09	85%	89%	4%	-0.08	-0.03
79244	1	194	0.57	0.57	0.00	85%	85%	0%	0.07	-0.12
80399	1	415	0.87	0.93	0.06	94%	97%	3%	0.04	0.00
81260	1	181	0.82	0.82	0.00	92%	92%	0%	0.03	-0.08
73863	2	187	0.69	0.74	0.05	73%	79%	6%	0.06	0.10
68311	2	207	0.85	0.88	0.04	82%	86%	4%	-0.01	-0.03
Avg.			0.75	0.80	0.05	86%	89%	3%	0.00	-0.03
Science										
70928	2	563	0.91	0.90	-0.01	90%	89%	0%	0.01	-0.06
70937	2	372	0.84	0.80	-0.04	89%	86%	-3%	0.06	-0.07
71344	2	240	0.88	0.87	-0.01	87%	85%	-2%	-0.01	-0.05
60001	2	259	0.81	0.82	0.00	80%	79%	0%	-0.05	-0.03
74531	2	335	0.92	0.92	0.00	89%	90%	2%	-0.03	-0.04
Avg.			0.87	0.86	-0.01	87%	86%	-1%	0.00	-0.05
Social Studies										
55826	2	358	0.56	0.57	0.01	66%	65%	-1%	0.05	-0.06
72841	2	770	0.74	0.72	-0.02	72%	69%	-4%	0.03	-0.08
72436	2	844	0.89	0.86	-0.03	89%	88%	-1%	0.00	-0.05
72439	2	309	0.78	0.80	0.02	77%	75%	-2%	0.00	-0.15
Avg.			0.75	0.74	-0.01	76%	74%	-2%	0.02	-0.08

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table 7. Phase 1 Percentage of STAAR SCRs Passing Performance Thresholds

Subject	N Items	QWK	Exact Agreement	SMD	Combined
RLA	10	100%	100%	90%	90%
Science	5	100%	100%	100%	100%
Social Studies	4	100%	100%	100%	100%

Phase 1: STAAR ECRs

Recall that ECR reported item scores are double the rubric score values for each dimension and represent the sum of rater scores. Recall also that the field-test programming sample sizes were smaller than what is typically used in engine programming.

Engine QWK values were greater than 0.70 for all items, and SMD values had a magnitude less than 0.15 except for item 55391 (Table 9). Aside from item 55391, the QWK values ranged from 0.73 to 0.89 for conventions and from 0.73 to 0.91 for ideas. Again ignoring item 55391, SMD values ranged from -0.04 to 0.06 for conventions and -0.03 to 0.07 for ideas. For RLA ECRs,

87.5% met the three thresholds (Table 10). Engine performance for each model appears in Appendix B: STAAR ECR Individual Model Performance Tables B1 and B2.

Table 8. Phase 1 AS Performance on STAAR ECRs

Item ID	N	QWK	Exact Agr.	Adj. Agr.	Non-Adj. Agr.	SMD	Mean		SD	
							HS	AS	HS	AS
Conventions										
55391	129	0.62	59%	34%	7%	0.28	0.71	0.47	0.95	0.84
12632	155	0.73	54%	36%	10%	0.05	1.16	1.10	1.28	1.14
12638	160	0.76	48%	45%	8%	0.03	1.52	1.48	1.36	1.19
12666	173	0.77	54%	31%	15%	0.06	1.43	1.35	1.48	1.39
73118	204	0.89	68%	26%	6%	-0.02	1.35	1.38	1.49	1.47
73991	187	0.89	64%	30%	6%	-0.04	1.82	1.89	1.58	1.56
68583	198	0.84	58%	36%	6%	-0.03	1.96	2.00	1.48	1.39
68776	173	0.77	58%	27%	15%	0.02	2.35	2.32	1.66	1.68
Avg.		0.78	58%	33%	9%	0.04				
Ideas										
55391	129	0.70	53%	36%	11%	0.12	1.29	1.14	1.28	1.13
12632	155	0.77	47%	41%	12%	0.07	1.66	1.56	1.56	1.38
12638	160	0.73	38%	48%	15%	0.02	2.05	2.02	1.57	1.45
12666	173	0.79	45%	41%	14%	-0.03	2.22	2.27	1.67	1.76
73118	204	0.91	67%	27%	6%	0.05	1.60	1.52	1.73	1.70
73991	187	0.89	53%	39%	8%	-0.01	1.99	2.01	1.83	1.76
68583	198	0.89	53%	38%	9%	0.03	2.36	2.31	1.89	1.89
68776	173	0.84	51%	35%	13%	0.02	2.18	2.14	1.71	1.77
Avg.		0.81	51%	38%	11%	0.03				

Note. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table 9. Phase 1 Percentage of STAAR ECRs Passing Performance Thresholds

N Items	Dimension	QWK	Exact Agreement	SMD	Combined
8	Conventions	87.5%	N.A.	87.5%	87.5%
8	Ideas	87.5%	N.A.	100%	87.5%

Phase 2: Operational Sample Scoring Categories Using FT-Programmed Models

After engine programming and validation on the field-test data, models were deployed to scoring environments and the operational subsamples of responses were scored by the engine. Following scoring, responses were categorized into those responses receiving engine condition codes, verification samples (15% determined randomly by day), low confidence (threshold less than 10),

and the remaining responses (AS). The total N count of scored responses and the percentage in each category are presented in Table 11.

The percentage of engine condition codes will vary depending upon the condition codes and thresholds used. Thus, the percentages displayed for STAAR should be interpreted with caution as the condition codes are provisional and yet to be finalized. Note also that a subset of the condition codes will be routed for human scoring in an operational setting.

The low confidence percentages for items should be around 10 since the percentile threshold was set to 10, which is the case for many items. However, there are thirteen STAAR items for which the low confidence percentages are more than 2% above or below 10%. This is likely due to a mismatch between the field-test sample responses and the operational sample responses, low N counts for engine programming, and distributional elements of the field-test sample. Items with the percentages outside of +/- 2% of expected low confidence percentage are underlined in the table.

Table 10. Phase 2 Percentage of Responses in the Four Categories for STAAR Items

Item ID	Subject	Type	N	Category			
				Machine CC	Verification	Low Conf.	AS
79024	RLA	SCR	88706	8%	14%	8%	71%
78742	RLA	SCR	90764	3%	15%	12%	71%
80822	RLA	SCR	92933	3%	15%	6%	76%
80104	RLA	SCR	97457	3%	15%	9%	73%
81164	RLA	SCR	99762	2%	15%	11%	71%
79244	RLA	SCR	102182	2%	15%	10%	73%
80399	RLA	SCR	128080	4%	14%	10%	72%
81260	RLA	SCR	116384	4%	14%	7%	75%
73863	RLA	SCR	102223	4%	14%	19%	63%
68311	RLA	SCR	128138	6%	14%	10%	69%
70928	Biology	SCR	113888	14%	13%	9%	64%
70937	Biology	SCR	114951	3%	14%	9%	73%
71344	Science	SCR	97024	0%	15%	6%	79%
60001	Science	SCR	101351	8%	14%	8%	69%
74531	Science	SCR	101511	13%	13%	6%	68%
55826	U.S. History	SCR	93917	4%	14%	11%	70%
72841	U.S. History	SCR	93969	6%	14%	7%	73%
72436	Social Studies	SCR	102972	6%	14%	6%	74%
72439	Social Studies	SCR	102887	10%	14%	6%	70%
55391	RLA	ECR	88772	14%	13%	14%	58%
12632	RLA	ECR	90858	8%	14%	9%	69%
12638	RLA	ECR	92935	4%	14%	16%	65%
12666	RLA	ECR	97421	5%	14%	9%	72%
73118	RLA	ECR	99679	4%	14%	10%	72%

Item ID	Subject	Type	N	Category			
				Machine CC	Verification	Low Conf.	AS
73991	RLA	ECR	101908	5%	14%	4%	76%
68583	RLA	ECR	127445	7%	14%	6%	73%
68776	RLA	ECR	115780	5%	14%	8%	72%

Note. CC = Condition Code.

Phase 3: Field-Test Programmed Model Performance on Operational Verification Data

Across item types, the automated scoring engine failed the performance criteria for many items on the operational verification data. Specifically, only 60% of STAAR RLA SCRs and 37.5% of ECRs met the performance criteria, and no science or social studies SCRs met the performance criteria.

Phase 3: STAAR SCRs

The automated scoring engine displayed generally lower QWK and exact agreement with the final, resolved score compared to the two human raters (Table 12). For some items, the engine-human QWK or exact agreement was much lower, such as for items 81164, 60001, 72436, and 72439. The SMD values tended to be negative across items indicating that the automated scoring engine assigned higher mean scores compared to the final, resolved scores. For some items, the SMD magnitude was quite large, exceeding 0.30 in magnitude (81164, 79244, 60001, 55826, 72841, 72436, and 72439). Only 60% of the RLA SCRs met the three thresholds in the verification data, and none of the science and social studies SRCs met the thresholds (Table 13).

Table 11. Phase 3 AS Performance on STAAR SCRs

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
RLA										
79024	1	12062	0.85	0.83	-0.03	93%	92%	-1%	0.00	-0.03
78742	1	13033	0.84	0.71	-0.12	92%	86%	-6%	-0.01	-0.19
80822	1	13440	0.74	0.67	-0.07	89%	86%	-3%	0.01	-0.04
80104	1	14078	0.78	0.80	0.01	89%	90%	1%	-0.02	-0.01
81164	1	14537	0.74	0.42	-0.32	87%	71%	-16%	-0.03	-0.33
79244	1	14893	0.80	0.65	-0.16	91%	84%	-7%	0.00	0.30
80399	1	18459	0.88	0.84	-0.05	94%	92%	-2%	0.00	-0.12
81260	1	16829	0.86	0.82	-0.04	93%	91%	-2%	-0.01	-0.05
73863	2	14711	0.71	0.71	0.00	76%	73%	-3%	-0.01	0.14
68311	2	18014	0.86	0.81	-0.05	83%	78%	-5%	0.00	-0.19
Avg			0.81	0.72	-0.08	89%	84%	-5%	-0.01	-0.05
Science										
70928	2	14653	0.89	0.82	-0.07	91%	81%	-9%	-0.02	-0.18
70937	2	16607	0.85	0.73	-0.13	88%	82%	-6%	0.01	-0.20

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
71344	2	14350	0.91	0.85	-0.06	88%	81%	-7%	0.00	0.01
60001	2	13846	0.82	0.65	-0.17	82%	64%	-18%	0.01	-0.31
74531	2	13201	0.97	0.92	-0.05	97%	91%	-5%	0.00	-0.01
Avg			0.89	0.79	-0.10	89%	80%	-9%	0.00	-0.14
Social Studies										
55826	2	13435	0.62	0.56	-0.06	62%	58%	-4%	0.01	0.30
72841	2	13215	0.55	0.56	0.00	62%	59%	-2%	0.01	-0.40
72436	2	14317	0.82	0.69	-0.14	80%	69%	-11%	-0.01	-0.41
72439	2	13879	0.81	0.65	-0.17	80%	61%	-19%	-0.02	-0.48
Avg			0.70	0.61	-0.09	71%	62%	-9%	0.00	-0.25

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold. STAAR SCRs had a 25% second read, so the human agreements were based upon a smaller sample.

Table 12. Phase 3 Percentage of STAAR SCRs Passing Performance Thresholds

Subject	N Items	QWK	Exact Agreement	SMD	Combined
RLA	10	70%	60%	60%	60%
Science	5	60%	0%	40%	0%
Social Studies	4	50%	50%	0%	0%

‘

Phase 3: STAAR ECRs

The automated scoring engine had QWK agreements exceeding 0.70 in the conventions and ideas dimensions except for items 55391 and 12632. Non-adjacent agreements were quite high for some items, exceeding 15% for six items and dimensions. The SMD magnitudes exceeded 0.15 for eight items (Table 14). SMD magnitudes exceeded 0.3 for four items and dimensions. As with the STAAR SCRs, the SMD were mostly negative; the engine assigned higher mean scores compared to the final, resolved score. Fifty percent of the items met all criteria for the conventions dimension and 37.5% met all criteria for the ideas dimension (Table 15). Engine performance for each model appears in Appendix B: STAAR ECR Individual Model Performance Tables B3 and B4.

Table 13. Phase 3 AS Performance on STAAR ECRs

Item ID	N	QWK	Exact Agr.	Adj. Agr.	Non-Adj. Agr.	SMD	Mean		SD	
							HS	AS	HS	AS
Conventions										
55391	11083	0.65	46%	36%	18%	0.07	1.36	1.27	1.32	1.38
12632	12337	0.65	45%	33%	22%	-0.44	1.35	2.01	1.49	1.50
12638	13142	0.77	47%	42%	11%	-0.17	1.80	2.04	1.40	1.39
12666	13636	0.82	54%	35%	11%	-0.08	1.93	2.06	1.55	1.55
73118	14248	0.83	55%	35%	10%	-0.09	2.03	2.17	1.50	1.64

Item ID	N	QWK	Exact Agr.	Adj. Agr.	Non-Adj. Agr.	SMD	Mean		SD	
							HS	AS	HS	AS
73991	14412	0.87	62%	31%	7%	-0.11	2.35	2.51	1.54	1.57
68583	17525	0.83	54%	39%	7%	-0.07	2.18	2.28	1.49	1.37
68776	16234	0.72	50%	31%	19%	-0.37	2.23	2.80	1.55	1.50
Ideas										
55391	11083	0.61	39%	38%	23%	-0.29	1.65	2.10	1.46	1.65
12632	12337	0.64	36%	36%	28%	-0.52	1.67	2.60	1.76	1.82
12638	13142	0.81	41%	41%	17%	-0.21	2.40	2.79	1.76	1.87
12666	13636	0.83	42%	39%	19%	-0.30	2.76	3.37	1.96	2.03
73118	14248	0.88	49%	38%	12%	0.02	2.67	2.63	1.86	2.13
73991	14412	0.91	53%	39%	8%	-0.01	2.90	2.92	2.02	2.08
68583	17525	0.89	51%	39%	10%	0.01	2.91	2.89	1.95	2.04
68776	16234	0.87	48%	38%	13%	-0.18	2.89	3.25	2.04	2.08

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table 14. Phase 3 Percentage of STAAR ECRs Passing Performance Thresholds

N Items	Dimension	QWK	Exact Agreement	SMD	Combined
8	Conventions	75%	N.A.	62.5%	50%
8	Ideas	75%	N.A.	37.5%	37.5%

Phase 4: Reprogrammed Model Performance on Verification Held-out Data

In phase 4, models were reprogrammed on the first 50% of the verification sample with 15% of that sample held out for model evaluation. This programming sample was chosen to mimic the fact that the engine performance would be monitored early in the window and, if not performing well, would be reprogrammed on a substantial portion of the verification sample that was considered reasonably representative of the set of testers throughout the administration. The automated scoring engine met the performance criteria for all items.

Phase 4: STAAR SCRs

The automated scoring engine had similar QWK and exact agreement with the final, resolved score relative to the two human raters (Table 16). The standardized mean differences between the engine and the final, resolved score ranged from -0.10 to 0.02 with many items exhibiting slightly negative SMD. In other words, the engine is assigning slightly higher mean scores for most items, but those values are close to zero. All STAAR SCRs met the three thresholds (Table 17).

Table 15. Phase 4 AS Performance on STAAR SCRs

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
RLA										
79024	1	896	0.90	0.88	-0.01	95%	94%	-1%	0.03	0.02
78742	1	980	0.86	0.84	-0.02	93%	92%	-1%	-0.01	-0.07
80822	1	1006	0.64	0.77	0.14	85%	90%	5%	-0.08	-0.01
80104	1	1060	0.79	0.83	0.04	90%	91%	2%	-0.05	-0.04
81164	1	1094	0.81	0.79	-0.02	91%	90%	-1%	-0.01	-0.06
79244	1	1116	0.86	0.85	-0.01	93%	93%	-1%	-0.04	-0.03
80399	1	1384	0.86	0.91	0.05	93%	96%	3%	0.04	-0.01
81260	1	1260	0.85	0.88	0.03	93%	94%	1%	0.00	-0.04
73863	2	1104	0.68	0.79	0.11	76%	83%	7%	-0.04	-0.05
68311	2	1351	0.85	0.89	0.05	82%	87%	5%	-0.08	0.00
Avg			0.81	0.84	0.03	89%	91%	2%	-0.02	-0.03
Science										
70928	2	1100	0.89	0.92	0.02	90%	93%	3%	-0.02	0.02
70937	2	1243	0.88	0.83	-0.05	90%	87%	-4%	0.00	-0.06
71344	2	1034	0.91	0.91	0.00	87%	87%	0%	-0.01	0.00
60001	2	1040	0.84	0.83	-0.01	84%	82%	-2%	0.01	-0.02
74531	2	986	0.96	0.98	0.02	96%	97%	2%	0.02	-0.01
Avg			0.90	0.89	0.00	89%	89%	0%	0.00	-0.01
Social Studies										
55826	2	1006	0.54	0.68	0.14	53%	65%	12%	-0.09	-0.10
72841	2	992	0.57	0.63	0.06	64%	68%	4%	0.05	-0.02
72436	2	1073	0.84	0.83	-0.02	82%	81%	-1%	0.03	0.00
72439	2	1041	0.83	0.83	-0.01	82%	82%	0%	0.01	-0.01
Avg			0.70	0.74	0.04	70%	74%	4%	0.00	-0.03

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold. STAAR SCRs had a 25% second read, so the human agreements were based upon a smaller sample.

Table 16. Phase 4 Percentage of STAAR SCRs Passing Performance Thresholds

Subject	N Items	QWK	Exact Agreement	SMD	Combined
RLA	10	100%	100%	100%	100%
Science	5	100%	100%	100%	100%
Social Studies	4	100%	100%	100%	100%

Phase 4: STAAR ECRs

For STAAR ECRs, the automated scoring engine had QWK agreements higher than 0.7 for all item dimensions with most QWK agreement values exceeding 0.80 (Table 18). Exact agreement rates varied between 52% and 61% for conventions and between 48% and 54% for ideas. Non-

adjacent agreement values ranged from 6% to 11% for conventions and from 6% to 12% for ideas. The SMD between the engine and the final, resolved score ranged from -0.10 to 0.02 for conventions and from -0.05 to 0.04 for ideas. All STAAR ECRs met the two thresholds (Table 19). Engine performance for each model appears in Appendix B: STAAR ECR Individual Model Performance Tables B5 and B6.

Table 17. Phase 4 AS Performance on STAAR ECRs

Item ID	N	QWK	Exact Agr.	Adj. Agr.	Non-Adj. Agr.	SMD	Mean		SD	
							HS	AS	HS	AS
Conventions										
55391	793	0.79	54%	40%	6%	-0.01	1.38	1.40	1.31	1.19
12632	806	0.83	58%	33%	9%	0.02	1.37	1.35	1.49	1.42
12638	965	0.77	52%	38%	11%	-0.10	1.82	1.95	1.40	1.35
12666	975	0.85	58%	33%	9%	0.02	1.94	1.91	1.54	1.54
73118	1044	0.85	57%	36%	7%	-0.02	2.03	2.06	1.48	1.45
73991	1047	0.88	61%	33%	6%	0.01	2.39	2.37	1.54	1.54
68583	1286	0.83	52%	41%	7%	-0.09	2.20	2.34	1.48	1.39
68776	1206	0.81	52%	39%	9%	-0.04	2.23	2.29	1.53	1.38
Avg.		0.83	55%	37%	8%	-0.03				
Ideas										
55391	793	0.81	53%	40%	7%	0.01	1.72	1.70	1.48	1.39
12632	806	0.84	54%	34%	12%	0.04	1.70	1.63	1.78	1.71
12638	965	0.83	48%	41%	11%	-0.02	2.42	2.44	1.76	1.63
12666	975	0.89	50%	42%	8%	0.04	2.79	2.71	1.94	1.93
73118	1044	0.89	54%	39%	6%	-0.03	2.64	2.69	1.83	1.71
73991	1047	0.92	54%	40%	6%	-0.01	2.97	2.99	2.02	2.08
68583	1286	0.89	50%	42%	8%	-0.05	2.98	3.07	1.94	1.92
68776	1206	0.90	50%	41%	8%	0.00	2.94	2.94	2.04	2.00
Avg.		0.87	52%	40%	8%	0.00				

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table 18. Phase 4 Percentage of STAAR ECRs Passing Performance Thresholds

N Items	Dimension	QWK	Exact Agreement	SMD	Combined
8	Conventions	100%	N.A.	100%	100%
8	Ideas	100%	N.A.	100%	100%

Phase 5: Operational Sample Scoring Categories Using Reprogrammed Models

After engine reprogramming and validation, models were deployed to the scoring environment, and the remaining AS sample from Phase 2 was rescored using the reprogrammed models. Following scoring, responses were categorized into those responses receiving any new engine condition codes and two low confidence thresholds (5 and 10). The total N count of scored responses and the percentage in each category are presented in Table 20.

As expected, the percentage of engine condition codes was very small for STAAR items. The newly assigned condition codes were those that were dependent upon the new data used to program the engine (e.g., Unusual Vocabulary, and Non-Specific). The low confidence values should be around 5 or 10 depending upon the threshold. Because low confidence responses were removed from the AS sample in Phase 2 using the field-test programmed model, it is possible that fewer than the threshold are routed. However, across item types, the percentages were close to the threshold.

Table 19. Phase 5 Percentage of Responses in the Four Categories for STAAR Items

Item ID	Subject	Type	N	Category		
				Machine CC	5th Percentile LC	10th Percentile LC
79024	RLA	SCR	62667	0.24%	5%	10%
78742	RLA	SCR	64161	0.09%	3%	8%
80822	RLA	SCR	70820	0.04%	3%	8%
80104	RLA	SCR	71306	0.07%	3%	6%
81164	RLA	SCR	71230	0.02%	3%	7%
79244	RLA	SCR	74949	0.02%	7%	11%
80399	RLA	SCR	91808	0.05%	3%	6%
81260	RLA	SCR	87642	0.03%	2%	4%
73863	RLA	SCR	64320	0.01%	6%	11%
68311	RLA	SCR	88851	0.01%	4%	9%
70928	Biology	SCR	73176	0.08%	4%	9%
70937	Biology	SCR	83907	0.05%	4%	8%
71344	Science	SCR	82469	0.63%	5%	10%
60001	Science	SCR	70309	0.08%	5%	9%
74531	Science	SCR	69044	0.08%	5%	11%
55826	U.S. History	SCR	66006	0.03%	5%	11%
72841	U.S. History	SCR	68235	0.02%	5%	9%
72436	Social Studies	SCR	75875	0.04%	4%	9%
72439	Social Studies	SCR	72328	0.01%	4%	10%
55391	RLA	ECR	51810	0.6%	5%	9%
12632	RLA	ECR	62954	0.8%	4%	9%
12638	RLA	ECR	60268	0.0%	3%	7%
12666	RLA	ECR	70068	0.4%	5%	9%

Item ID	Subject	Type	N	Category		
				Machine CC	5th Percentile LC	10th Percentile LC
73118	RLA	ECR	71446	0.0%	5%	9%
73991	RLA	ECR	77701	0.0%	5%	9%
68583	RLA	ECR	93641	0.4%	5%	10%
68776	RLA	ECR	83881	0.0%	4%	7%

Note. CC = Condition Code. LC = Low Confidence.

Phase 6: Reprogrammed Model Performance on Operational AS Data

The reprogrammed automated scoring engine models met the performance criteria for all items on the rescored operational AS data. This was true also for the five student group categories (Male, Female, Black, Hispanic/Latino, and White).

Phase 6: STAAR SCRs

The automated scoring engine showed comparable QWK and exact agreement with the final, resolved score compared to the two human raters (Table 21). The average SMD between the engine and the final, resolved score ranged from -0.07 to 0.02. Most SMD values were negative indicating that the engine assigned slightly higher mean scores compared to the final, resolved score. All STAAR SCRs met the three performance criteria overall and by student group (Table 22). The model performance by student group for each item is in Appendix C: Subgroup Item-Level Results Tables C1, C2, and C3.

Table 20. Phase 6 AS Performance on STAAR SCRs

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
RLA										
79024	1	61555	0.87	0.89	0.02	94%	95%	1%	0.00	0.02
78742	1	63272	0.87	0.88	0.01	94%	94%	1%	0.00	-0.03
80822	1	70352	0.76	0.80	0.05	90%	92%	2%	-0.01	0.00
80104	1	70524	0.81	0.87	0.06	91%	94%	3%	0.00	-0.02
81164	1	70871	0.77	0.83	0.06	88%	91%	3%	0.00	-0.05
79244	1	74656	0.76	0.82	0.06	89%	92%	3%	0.00	-0.01
80399	1	91672	0.91	0.93	0.02	96%	97%	1%	0.00	-0.01
81260	1	87586	0.89	0.92	0.03	95%	96%	1%	0.00	-0.01
73863	2	63949	0.73	0.79	0.06	77%	82%	5%	0.00	-0.02
68311	2	88792	0.88	0.91	0.03	85%	88%	3%	0.00	-0.01
Avg			0.83	0.86	0.04	90%	92%	2%	0.00	-0.01
Science										
70928	2	72639	0.91	0.92	0.01	92%	93%	1%	0.00	0.01
70937	2	83660	0.86	0.83	-0.02	89%	88%	-1%	0.00	-0.07

Item ID	Score Point	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
71344	2	78607	0.91	0.90	0.00	87%	87%	0%	0.00	0.00
60001	2	69840	0.83	0.84	0.00	83%	82%	0%	0.01	-0.06
74531	2	68633	0.98	0.97	0.00	97%	97%	0%	0.00	-0.01
Avg			0.90	0.89	0.00	90%	90%	0%	0.00	-0.03
Social Studies										
55826	2	65419	0.59	0.64	0.06	61%	64%	3%	-0.01	-0.07
72841	2	68073	0.58	0.65	0.07	63%	68%	6%	-0.01	-0.03
72436	2	75139	0.81	0.83	0.02	80%	82%	2%	0.00	0.00
72439	2	71925	0.84	0.85	0.01	82%	82%	0%	0.00	0.00
Avg			0.70	0.74	0.04	71%	74%	3%	0.00	-0.02

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold. STAAR SCRs had a 25% second read, so the human agreements were based upon a smaller sample.

Table 21. Phase 6 Percentage of STAAR SCRs Passing Performance Thresholds

Subgroup	QWK	Exact Agreement	SMD	Combined
RLA				
Overall	100%	100%	100%	100%
Females	100%	100%	100%	100%
Males	100%	100%	100%	100%
Black	100%	100%	100%	100%
Hispanic/Latino	100%	100%	100%	100%
White	100%	100%	100%	100%
Science				
Overall	100%	100%	100%	100%
Females	100%	100%	100%	100%
Males	100%	100%	100%	100%
Black	100%	100%	100%	100%
Hispanic/Latino	100%	100%	100%	100%
White	100%	100%	100%	100%
Social Studies				
Overall	100%	100%	100%	100%
Females	100%	100%	100%	100%
Males	100%	100%	100%	100%
Black	100%	100%	100%	100%
Hispanic/Latino	100%	100%	100%	100%
White	100%	100%	100%	100%

Phase 6: STAAR ECRs

The automated scoring engine had generally high QWK agreements for STAAR ECRs with all QWK values greater than 0.80 (Table 23). Exact agreement rates ranged from 52% to 62% for conventions and from 51% to 58% for ideas. Non-adjacent agreements were 10% or less for all items and dimensions except item 12632 in ideas. SMD magnitudes varied between -0.08 and 0.01 in conventions and between -0.04 and 0.03 in ideas. All items and dimensions met the two performance criteria overall and by student group (Table 24). Engine performance for each model appears in Appendix B: STAAR ECR Individual Model Performance Tables B7 and B8. Performance by student group for each item is in Appendix C: Subgroup Item-Level Results Table C4.

Table 22. Phase 6 AS Performance on STAAR ECRs

Item ID	N	QWK	Exact Agr.	Adj. Agr.	Non-Adj. Agr.	SMD	Mean		SD	
							HS	AS	HS	AS
Conventions										
55391	47799	0.81	58%	36%	6%	-0.01	1.29	1.30	1.31	1.24
12632	54564	0.82	60%	31%	10%	0.00	1.35	1.34	1.50	1.44
12638	57547	0.80	52%	39%	9%	-0.08	1.60	1.71	1.40	1.36
12666	65784	0.86	58%	34%	8%	0.01	2.00	1.99	1.59	1.59
73118	69048	0.86	58%	35%	6%	-0.02	2.07	2.10	1.53	1.52
73991	74461	0.88	62%	32%	6%	0.01	2.44	2.42	1.55	1.54
68583	89627	0.85	55%	38%	6%	-0.08	2.23	2.35	1.50	1.45
68776	81190	0.83	55%	37%	8%	-0.05	2.33	2.41	1.56	1.43
Ideas										
55391	47799	0.83	58%	35%	7%	0.00	1.60	1.60	1.47	1.42
12632	54564	0.84	55%	33%	12%	0.03	1.69	1.63	1.79	1.73
12638	57547	0.87	51%	40%	9%	0.00	2.16	2.16	1.78	1.72
12666	65784	0.90	52%	40%	8%	0.02	2.85	2.80	2.02	2.02
73118	69048	0.90	54%	40%	6%	-0.03	2.74	2.80	1.91	1.83
73991	74461	0.92	56%	38%	6%	0.00	3.02	3.03	2.04	2.07
68583	89627	0.91	54%	39%	7%	-0.04	3.01	3.10	1.98	2.01
68776	81190	0.91	52%	40%	8%	-0.01	3.00	3.03	2.07	2.07

Note. H1H2 refers to human-human agreement. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table 23. Phase 6 Percentage of STAAR ECRs Passing Performance Thresholds

Subgroup	QWK	Exact Agreement	SMD	Combined
Conventions				
Overall	100%	N.A.	100%	100%
Females	100%	N.A.	100%	100%
Males	100%	N.A.	100%	100%
Black	100%	N.A.	100%	100%

Subgroup	QWK	Exact Agreement	SMD	Combined
Hispanic/Latino	100%	N.A.	100%	100%
White	100%	N.A.	100%	100%
Ideas				
Overall	100%	N.A.	100%	100%
Females	100%	N.A.	100%	100%
Males	100%	N.A.	100%	100%
Black	100%	N.A.	100%	100%
Hispanic/Latino	100%	N.A.	100%	100%
White	100%	N.A.	100%	100%

‘

Conclusion and Next Steps

This study built upon prior studies examining the performance of automated scoring on the STAAR SCRs and ECRs. The purpose of the study was to examine the robustness of the models programmed with field-test data on the operational responses and scores using spring 2023 items and data. We expect that the results of the study will generalize to future administrations beginning in December 2023.

The study found that the automated scoring engine met all or most performance criteria on each of the field-test programmed and operationally programmed held-out validation samples. However, many models programmed with the field-test data did not perform well on the operational data. Specifically, only 60% of STAAR RLA SCRs and 37.5% of STAAR ECRs met the performance criteria while none of the STAAR science or social studies SCRs met the performance criteria.

Models programmed on the sample of the first 50% of the operational data in the program met performance criteria on the non-routed operational data for all students and for the five student groups evaluated (Male, Female, Black, Hispanic/Latino, and White).

These results suggest that, in future administrations, we should expect that most field-test programmed models will not meet the performance criteria and that models will need to be reprogrammed on the operational data. Moving forward, we recommend that all models for all items be reprogrammed on the operational data, regardless of performance. We recommend that field-test programmed models be used initially for operational scoring while routing responses with condition codes, low confidence, and a random verification sample to human scoring. We also recommend that the automated score not be used in the human scoring process; rather, any response routed for hand-scoring use spring 2023 hand-scoring rules to assign scores. This approach will better support engine reprogramming on the operational data, allowing for a direct comparison of the engine to humans. During the operational window, we recommend new models be reprogrammed on the operational data when a sufficient sample size is determined (e.g., minimum of 3,000 to ensure sufficiently large held-out validation samples are used to evaluate performance). Once deployed, we recommend that all responses be rescored using the newly programmed model with any newly determined condition codes or low confidence responses routed for human scoring. Finally, we recommend that any scores assigned during human scoring

be considered the score of record; all other responses receive the score assigned by the reprogrammed model.

Additional work is needed to continue to refine the condition codes and thresholds used. These will likely vary by item type and grade and will undergo review and analysis to ensure alignment to the scoring rubric and human-assigned condition codes.

References

Cambium Assessment, Inc. (2021). *Technical Report: Automated Scoring Performance on the NAEP Automated Scoring Challenge: Item Specific Models.* Retrieved from https://www.cambiumassessment.com/-/media/project/cambium/corporate/pdfs/naep_technical_report_item_specific_models_final.pdf.

Clark, K., Luong, M-T, Le, Q., & Manning, C (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators.* Paper presented at the Seventh International Conference on Learning Representations. arXiv. https://doi.org/10.48550/arXiv.2003.10555.

Deerwester, S. Dumais, S.T., & Landauer, T.K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6),* 391–407.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal archive, 12*, 23-38.

Jiang, Z. W., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). *Convbert: Improving bert with span-based dynamic convolution.* Paper presented at the 34th Confernce on Neural Information Processing Systems (NeurIPS), Vancouver, CA. arXiv. https://doi.org/10.48550/arXiv.2008.02496.

PARCC (2015, March 9). *Research Results of PARCC Automated Scoring Proof of Concept Study.* Retrieved from: http://www.parcconline.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need.* Paper pesented at the 31st Conference on Neural Information Processing Systems.

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1),* 2-13.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*, 239-263.

Appendix A: Field-Test and Operational Score Distributions

Table A1. SAFT and Operational Performance on STAAR RLA SCR Items

Item ID	Admin	N	Mean	SD	SMD	Percent in Score Category			
						0	1	2	CC
79024	SAFT	1436	0.50	0.50	0.06	50%	50%		4%
79024	OP	512545	0.54	0.50		46%	54%		4%
78742	SAFT	1289	0.44	0.50	0.08	56%	44%		4%
78742	OP	408888	0.48	0.50		52%	48%		3%
80822	SAFT	2570	0.60	0.49	0.13	40%	60%		1%
80822	OP	363053	0.66	0.47		34%	66%		2%
80104	SAFT	2431	0.44	0.50	0.12	56%	44%		2%
80104	OP	354824	0.50	0.50		50%	50%		3%
81164	SAFT	1334	0.46	0.50	0.05	54%	46%		1%
81164	OP	408725	0.49	0.50		51%	49%		2%
79244	SAFT	2578	0.21	0.41	0.41	79%	21%		2%
79244	OP	389822	0.40	0.49		60%	40%		2%
80399	SAFT	2871	0.61	0.49	-0.20	39%	61%		3%
80399	OP	512313	0.51	0.50		49%	51%		2%
81260	SAFT	1182	0.57	0.50	-0.07	43%	57%		5%
81260	OP	371728	0.54	0.50		46%	54%		1%
73863	SAFT	1240	0.90	0.65	0.57	26%	57%	17%	3%
73863	OP	399038	1.30	0.73		16%	38%	46%	2%
68311	SAFT	1279	1.08	0.77	0.06	26%	40%	34%	2%
68311	OP	465531	1.12	0.81		28%	32%	40%	1%

Note. SMD values underlined when they exceed 0.15.

Table A2. SAFT and Operational Performance on STAAR Science and Social Studies SCR Items

Item ID	Admin	Subject	N	Mean	SD	SMD	Percent in Score Category			
							0	1	2	CC
70928	SAFT	Biology	4339	0.66	0.75	-0.07	51%	33%	17%	7%
70928	OP	Biology	455242	0.61	0.66		50%	40%	10%	6%
70937	SAFT	Biology	2542	1.47	0.67	-0.15	10%	33%	57%	1%
70937	OP	Biology	459605	1.37	0.74		15%	32%	52%	2%
71344	SAFT	Science	1656	0.54	0.78	0.20	64%	18%	18%	1%
71344	OP	Science	388092	0.71	0.85		56%	18%	26%	1%
60001	SAFT	Science	1978	0.81	0.78	-0.17	42%	36%	22%	2%

Item ID	Admin	Subject	N	Mean	SD	SMD	Percent in Score Category			
							0	1	2	CC
60001	OP	Science	405398	0.68	0.74		49%	35%	16%	4%
71344	SAFT	Science	2742	1.02	0.86	0.21	36%	26%	38%	2%
71344	OP	Science	406040	1.21	0.86		29%	21%	50%	2%
72436	SAFT	Social Studies	6099	1.32	0.80	-0.07	21%	26%	53%	4%
72436	OP	Social Studies	411882	1.26	0.83		25%	24%	51%	4%
72439	SAFT	Social Studies	2493	0.82	0.85	-0.27	47%	25%	28%	11%
72439	OP	Social Studies	411544	0.59	0.78		59%	23%	18%	4%
55826	SAFT	U.S. History	2474	0.69	0.71	0.34	45%	41%	14%	4%
55826	OP	U.S. History	375359	0.94	0.78		34%	38%	28%	3%
72841	SAFT	U.S. History	6510	0.84	0.80	0.40	41%	34%	25%	12%
72841	OP	U.S. History	375457	1.14	0.74		21%	43%	36%	3%

Note. SMD values underlined when they exceed 0.15.

Table A3. SAFT and Operational Performance on STAAR RLA ECR Items—Conventions

Item ID	Admin	N	Mean	SD	SMD	Percent in Score Category					
						0	1	2	3	4	CC
55391	SAFT	1259	0.54	0.91	0.54	66%	20%	10%	2%	2%	12%
55391	OP	355087	1.14	1.30		48%	14%	22%	9%	7%	9%
12632	SAFT	1289	0.97	1.26	0.17	54%	15%	17%	7%	7%	11%
12632	OP	363427	1.21	1.46		51%	13%	13%	11%	12%	5%
12638	SAFT	1165	1.45	1.36	0.17	35%	20%	21%	14%	10%	4%
12638	OP	371737	1.69	1.42		31%	15%	24%	16%	15%	3%
12666	SAFT	1291	1.32	1.47	0.32	47%	12%	16%	13%	12%	5%
12666	OP	389680	1.81	1.58		34%	13%	15%	16%	22%	4%
73118	SAFT	1420	1.30	1.49	0.43	49%	10%	18%	10%	14%	3%
73118	OP	398709	1.95	1.52		27%	14%	19%	17%	23%	3%
73991	SAFT	1353	1.65	1.57	0.36	39%	11%	17%	12%	21%	5%
73991	OP	407627	2.22	1.60		26%	10%	15%	17%	33%	4%
68583	SAFT	1400	1.83	1.48	0.11	30%	12%	24%	16%	19%	5%
68583	OP	509294	2.01	1.55		28%	10%	19%	18%	24%	7%
68776	SAFT	1267	2.10	1.71	-0.01	32%	9%	10%	14%	35%	7%
68776	OP	462885	2.09	1.60		28%	10%	15%	18%	29%	5%

Note. SMD values underlined when they exceed 0.15.

Table A4. SAFT and Operational Performance on STAAR RLA ECR Items–Ideas

Item ID	Admin	N	Mean	SD	SMD	Percent in Score Category						
						0	1	2	3	4	5	6
55391	SAFT	1259	0.90	1.18	0.37	53%	18%	21%	4%	3%	1%	0%

Item ID	Admin	N	Mean	SD	SMD	Percent in Score Category						
						0	1	2	3	4	5	6
55391	OP	355087	1.39	1.46		42%	10%	29%	9%	7%	2%	1%
12632	SAFT	1289	1.38	1.54	0.08	41%	17%	22%	8%	7%	2%	2%
12632	OP	363427	1.51	1.75		46%	12%	13%	10%	11%	5%	2%
12638	SAFT	1165	1.84	1.58	0.24	26%	18%	27%	13%	10%	4%	3%
12638	OP	371737	2.25	1.79		25%	12%	23%	14%	15%	7%	5%
12666	SAFT	1291	1.94	1.74	0.34	29%	15%	23%	11%	13%	5%	4%
12666	OP	389680	2.58	2.02		23%	11%	18%	12%	15%	9%	11%
73118	SAFT	1420	1.55	1.72	0.56	44%	9%	25%	7%	10%	3%	4%
73118	OP	398709	2.56	1.89		19%	11%	24%	12%	15%	9%	9%
73991	SAFT	1353	1.85	1.85	0.45	35%	12%	24%	8%	10%	5%	6%
73991	OP	407627	2.73	2.07		23%	9%	17%	11%	15%	12%	13%
68583	SAFT	1400	2.21	1.92	0.24	29%	12%	19%	11%	16%	7%	7%
68583	OP	509199	2.68	2.04		24%	9%	17%	10%	18%	12%	11%
68776	SAFT	1267	1.99	1.76	0.36	27%	13%	30%	9%	9%	7%	5%
68776	OP	463011	2.69	2.10		24%	10%	16%	11%	15%	12%	13%

Note. SMD values underlined when they exceed 0.15.

Appendix B: STAAR ECR Individual Model Performance

Phase 1: STAAR ECR Items

Models showed slightly lower QWK and exact agreement with the final, resolved score relative to the two human raters (Table B1). Item 55391 had much worse QWK performance for both models in ideas, and the SMD for model 1 in conventions was large (0.41). Both models met all criteria for 63% of the conventions dimensions; model 1 met all criteria for 50% of the ideas dimension, and model 2 met all criteria for 75% of the ideas dimension (Table B2). Most failures to meet the criteria in conventions were for the SMD threshold, and most failures in ideas were for exact agreement. Almost all items met the QWK criterion except item 55391.

Table B1. Phase 1 Model Performance on STAAR ECR Items

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
Conventions										
55391	1	129	0.48	0.42	-0.07	69%	74%	5%	0.08	0.41

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
55391	2	129	0.48	0.49	0.01	69%	72%	3%	0.08	0.12
12632	1	155	0.61	0.58	-0.03	70%	63%	-7%	-0.07	0.18
12632	2	155	0.61	0.62	0.01	70%	67%	-3%	-0.07	0.06
12638	1	160	0.53	0.52	-0.01	58%	61%	2%	0.02	0.23
12638	2	160	0.53	0.66	0.13	58%	62%	4%	0.02	-0.09
12666	1	173	0.73	0.70	-0.03	72%	69%	-3%	-0.09	0.12
12666	2	173	0.73	0.67	-0.06	72%	66%	-6%	-0.09	0.03
73118	1	204	0.83	0.78	-0.05	79%	75%	-4%	-0.03	0.14
73118	2	204	0.83	0.81	-0.02	79%	77%	-2%	-0.03	-0.12
73991	1	187	0.80	0.79	-0.01	74%	74%	-1%	0.01	0.15
73991	2	187	0.80	0.80	0.00	74%	76%	2%	0.01	-0.19
68583	1	198	0.73	0.71	-0.02	71%	67%	-4%	0.01	0.08
68583	2	198	0.73	0.77	0.04	71%	73%	2%	0.01	-0.19
68776	1	173	0.60	0.62	0.02	67%	69%	2%	0.09	0.03
68776	2	173	0.60	0.66	0.07	67%	67%	0%	0.09	0.07
Ideas										
55391	1	129	0.69	0.55	-0.14	71%	65%	-5%	0.07	0.13
55391	2	129	0.69	0.56	-0.14	71%	64%	-6%	0.07	0.11
12632	1	155	0.64	0.59	-0.04	61%	57%	-4%	0.01	0.18
12632	2	155	0.64	0.66	0.02	61%	66%	5%	0.01	0.00
12638	1	160	0.54	0.55	0.01	56%	52%	-4%	-0.08	0.08
12638	2	160	0.54	0.62	0.08	56%	55%	-1%	-0.08	-0.01
12666	1	173	0.72	0.69	-0.04	64%	60%	-3%	-0.06	0.06
12666	2	173	0.72	0.73	0.00	64%	63%	-1%	-0.06	-0.12
73118	1	204	0.90	0.83	-0.07	84%	74%	-10%	-0.01	0.03
73118	2	204	0.90	0.86	-0.04	84%	79%	-4%	-0.01	0.03
73991	1	187	0.85	0.78	-0.07	73%	64%	-9%	0.01	0.09
73991	2	187	0.85	0.84	-0.01	73%	74%	1%	0.01	-0.10
68583	1	198	0.82	0.81	-0.01	67%	68%	1%	-0.03	0.08
68583	2	198	0.82	0.83	0.01	67%	69%	3%	-0.03	-0.02
68776	1	173	0.75	0.73	-0.01	65%	62%	-2%	0.01	0.09
68776	2	173	0.75	0.74	-0.01	65%	58%	-6%	0.01	-0.03

Note. HSAS refers to the agreement of engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table B2. Phase 1 Percentage of Models Passing Performance Thresholds

Model	QWK	Exact Agreement	SMD	Combined
Conventions				
1	100%	88%	63%	63%
2	100%	88%	75%	63%

Model	QWK	Exact Agreement	SMD	Combined
Ideas				
1	88%	63%	88%	50%
2	88%	75%	100%	75%

Phase 3: STAAR ECR Items

Both models showed slightly lower QWK and exact agreement with the final, resolved score relative to the two human raters with large differences for items 55391 and 12632 (Table B3). In conventions, three items had large SMD magnitudes for either model (12632, 12638, and 68776). In ideas, five items had large SMD magnitudes for either model (55391, 12632, 12638, 12666, and 68776).

Model 1 met all the combined criteria for 63% of the items in both conventions and ideas. Models met all the criteria for 25% of items in conventions and 38% in ideas (Table B4). Most model 2 failures were for SMD violations.

Table B3. Phase 3 Model Performance on STAAR ECR Items

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
Conventions										
55391	1	11083	0.68	0.58	-0.10	71%	62%	-9%	0.01	0.11
55391	2	11083	0.68	0.54	-0.14	71%	60%	-11%	0.01	0.02
12632	1	12337	0.67	0.59	-0.08	70%	61%	-9%	-0.01	-0.37
12632	2	12337	0.67	0.56	-0.11	70%	58%	-12%	-0.01	-0.48
12638	1	13142	0.65	0.64	-0.01	65%	63%	-2%	0.00	-0.02
12638	2	13142	0.65	0.64	-0.01	65%	60%	-5%	0.00	-0.31
12666	1	13636	0.69	0.70	0.00	66%	66%	0%	0.01	-0.04
12666	2	13636	0.69	0.72	0.03	66%	67%	1%	0.01	-0.12
73118	1	14248	0.69	0.73	0.03	66%	67%	1%	0.00	-0.01
73118	2	14248	0.69	0.74	0.05	66%	69%	3%	0.00	-0.17
73991	1	14412	0.74	0.78	0.04	70%	72%	2%	0.01	0.03
73991	2	14412	0.74	0.76	0.03	70%	72%	3%	0.01	-0.24
68583	1	17525	0.70	0.71	0.00	67%	66%	-1%	0.01	-0.01
68583	2	17525	0.70	0.71	0.00	67%	68%	0%	0.01	-0.12
68776	1	16234	0.68	0.58	-0.11	66%	61%	-5%	0.00	-0.41
68776	2	16234	0.68	0.64	-0.04	66%	64%	-2%	0.00	-0.28
Ideas										
55391	1	11083	0.74	0.56	-0.18	74%	54%	-20%	-0.01	-0.20
55391	2	11083	0.74	0.51	-0.23	74%	56%	-17%	-0.01	-0.29
12632	1	12337	0.72	0.60	-0.12	66%	50%	-16%	-0.01	-0.47

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
12632	2	12337	0.72	0.53	-0.18	66%	51%	-14%	-0.01	-0.44
12638	1	13142	0.72	0.71	-0.01	61%	57%	-4%	0.00	-0.09
12638	2	13142	0.72	0.69	-0.03	61%	55%	-6%	0.00	-0.29
12666	1	13636	0.78	0.78	0.00	62%	62%	-1%	0.00	-0.20
12666	2	13636	0.78	0.73	-0.06	62%	54%	-8%	0.00	-0.40
73118	1	14248	0.78	0.81	0.02	64%	63%	-1%	-0.01	-0.02
73118	2	14248	0.78	0.82	0.03	64%	65%	1%	-0.01	0.06
73991	1	14412	0.81	0.83	0.02	64%	65%	1%	-0.01	0.06
73991	2	14412	0.81	0.84	0.03	64%	68%	4%	-0.01	-0.10
68583	1	17525	0.80	0.81	0.02	64%	65%	0%	0.01	0.07
68583	2	17525	0.80	0.82	0.02	64%	66%	1%	0.01	-0.05
68776	1	16234	0.79	0.79	0.00	62%	61%	-1%	-0.01	-0.11
68776	2	16234	0.79	0.79	0.00	62%	60%	-2%	-0.01	-0.26

Note. HSAS refers to the agreement of engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table B4. Phase 3 Percentage of Models Passing Performance Thresholds

Model	QWK	Exact Agreement	SMD	Combined
Conventions				
1	88%	63%	75%	63%
2	75%	75%	38%	25%
Ideas				
1	75%	75%	63%	63%
2	75%	50%	38%	38%

Phase 4: STAAR ECR Items

Both models showed comparable QWK and exact agreement with the final, resolved score relative to the two human raters, except for item 55391 (Table B5). Model 1 all criteria for 75% of the Conventions dimensions and Model 2 met all criteria for 63% of the items, with all violations at the SMD criterion. For Ideas, 88% of the items met all criteria for Model 1 and 100% met the criteria for model 2 (Table B6).

Table B5. Phase 4 Model Performance on STAAR ECR Items

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
Conventions										
55391	1	793	0.66	0.60	-0.06	70%	64%	-6%	0.00	0.13
55391	2	793	0.66	0.71	0.05	70%	71%	1%	0.00	-0.15

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
12632	1	806	0.71	0.67	-0.04	71%	68%	-3%	-0.01	0.18
12632	2	806	0.71	0.74	0.02	71%	72%	0%	-0.01	-0.14
12638	1	965	0.66	0.65	-0.01	65%	66%	0%	0.00	0.01
12638	2	965	0.66	0.65	-0.01	65%	64%	-1%	0.00	-0.18
12666	1	975	0.69	0.74	0.05	65%	70%	5%	0.02	0.10
12666	2	975	0.69	0.75	0.06	65%	70%	5%	0.02	-0.03
73118	1	1044	0.70	0.72	0.02	65%	68%	3%	0.00	0.05
73118	2	1044	0.70	0.76	0.06	65%	71%	6%	0.00	-0.10
73991	1	1047	0.72	0.76	0.04	69%	71%	2%	0.03	0.02
73991	2	1047	0.72	0.79	0.07	69%	72%	3%	0.03	-0.01
68583	1	1286	0.72	0.69	-0.02	68%	65%	-3%	0.02	-0.02
68583	2	1286	0.72	0.73	0.02	68%	69%	0%	0.02	-0.16
68776	1	1206	0.64	0.67	0.03	64%	63%	-1%	-0.01	0.00
68776	2	1206	0.64	0.69	0.05	64%	67%	2%	-0.01	-0.10
Ideas										
55391	1	793	0.74	0.63	-0.11	72%	62%	-10%	-0.03	0.12
55391	2	793	0.74	0.76	0.02	72%	74%	2%	-0.03	-0.07
12632	1	806	0.73	0.73	0.00	67%	65%	-1%	-0.01	0.14
12632	2	806	0.73	0.76	0.02	67%	69%	3%	-0.01	-0.05
12638	1	965	0.72	0.71	-0.01	62%	61%	-1%	0.01	0.04
12638	2	965	0.72	0.76	0.04	62%	65%	3%	0.01	-0.04
12666	1	975	0.78	0.80	0.02	63%	65%	2%	-0.03	0.04
12666	2	975	0.78	0.83	0.05	63%	69%	6%	-0.03	0.06
73118	1	1044	0.79	0.81	0.02	65%	68%	3%	-0.01	0.05
73118	2	1044	0.79	0.81	0.03	65%	70%	5%	-0.01	-0.11
73991	1	1047	0.79	0.85	0.06	62%	69%	8%	0.00	0.01
73991	2	1047	0.79	0.85	0.06	62%	69%	8%	0.00	-0.07
68583	1	1286	0.81	0.81	0.01	65%	65%	0%	0.03	0.06
68583	2	1286	0.81	0.83	0.02	65%	66%	1%	0.03	-0.15
68776	1	1206	0.78	0.82	0.04	62%	66%	4%	0.02	0.04
68776	2	1206	0.78	0.83	0.05	62%	67%	6%	0.02	-0.08

Note. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table B6. Phase 4 Percentage of Models Passing Performance Thresholds

Model	QWK	Exact Agreement	SMD	Combined
Conventions				
1	100%	88%	88%	75%
2	100%	100%	63%	63%
Ideas				

Model	QWK	Exact Agreement	SMD	Combined
1	88%	88%	100%	88%
2	100%	100%	100%	100%

Phase 6: STAAR ECR Items

Both models showed comparable QWK and exact agreement with the final, resolved score relative to the two human raters except for item 55391 (Table B7). Model 1 met all criteria for 88% of the conventions dimensions, and model 2 met all criteria for 100% of the conventions dimension with all violations at the SMD criterion. For ideas, 88% of the items met all criteria for model 1, and 100% met the criteria for model 2 (Table B8).

Table B7. Phase 6 Model Performance on STAAR ECR Items

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
Conventions										
55391	1	47799	0.69	0.64	-0.05	72%	68%	-5%	0.00	0.12
55391	2	47799	0.69	0.72	0.03	72%	74%	2%	0.00	-0.14
12632	1	54564	0.69	0.66	-0.03	71%	69%	-3%	0.00	0.16
12632	2	54564	0.69	0.73	0.04	71%	72%	1%	0.00	-0.14
12638	1	57547	0.66	0.67	0.01	67%	65%	-1%	0.01	-0.01
12638	2	57547	0.66	0.67	0.01	67%	66%	-1%	0.01	-0.14
12666	1	65784	0.72	0.74	0.02	67%	69%	1%	0.00	0.06
12666	2	65784	0.72	0.78	0.06	67%	72%	5%	0.00	-0.04
73118	1	69048	0.71	0.74	0.03	67%	69%	2%	0.00	0.03
73118	2	69048	0.71	0.78	0.07	67%	73%	6%	0.00	-0.09
73991	1	74461	0.75	0.78	0.03	71%	73%	2%	0.00	0.00
73991	2	74461	0.75	0.81	0.05	71%	76%	5%	0.00	0.00
68583	1	89627	0.72	0.72	0.00	68%	67%	-2%	0.01	-0.03
68583	2	89627	0.72	0.76	0.04	68%	72%	3%	0.01	-0.15
68776	1	81190	0.70	0.72	0.02	67%	67%	-1%	0.00	-0.04
68776	2	81190	0.70	0.72	0.02	67%	69%	1%	0.00	-0.11
Ideas										
55391	1	47799	0.75	0.67	-0.08	75%	66%	-9%	0.00	0.09
55391	2	47799	0.75	0.76	0.01	75%	76%	1%	0.00	-0.09
12632	1	54564	0.73	0.71	-0.02	66%	63%	-3%	0.00	0.14
12632	2	54564	0.73	0.77	0.04	66%	69%	3%	0.00	-0.05
12638	1	57547	0.75	0.77	0.02	64%	65%	1%	0.00	0.01
12638	2	57547	0.75	0.79	0.04	64%	67%	3%	0.00	-0.01
12666	1	65784	0.80	0.82	0.03	63%	65%	3%	0.00	0.04
12666	2	65784	0.80	0.84	0.04	63%	67%	5%	0.00	0.01
73118	1	69048	0.79	0.82	0.03	64%	68%	4%	0.00	0.03

Item ID	Model	N	QWK			Exact Agreement			SMD	
			H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
73118	2	69048	0.79	0.83	0.03	64%	69%	5%	0.00	-0.11
73991	1	74461	0.82	0.85	0.03	64%	69%	5%	0.00	0.03
73991	2	74461	0.82	0.86	0.04	64%	70%	6%	0.00	-0.05
68583	1	89627	0.81	0.83	0.03	65%	68%	3%	0.00	0.04
68583	2	89627	0.81	0.84	0.03	65%	68%	4%	0.00	-0.13
68776	1	81190	0.80	0.83	0.03	63%	66%	3%	0.00	0.02
68776	2	81190	0.80	0.84	0.04	63%	67%	4%	0.00	-0.09

Note. HSAS refers to the agreement of the engine with the final, resolved score. A positive HSAS SMD means the HS mean is higher than the AS mean. Underlined values indicate violation of a performance threshold.

Table B8. Phase 6 Percentage of Models Passing Performance Thresholds

Model	QWK	Exact Agreement	SMD	Combined
Conventions				
1	100%	100%	88%	88%
2	100%	100%	100%	100%
Ideas				
1	100%	88%	100%	88%
2	100%	100%	100%	100%

Appendix C: Subgroup Item-Level Results

Table C1. STAAR RLA SCR Subgroup Performance

Item ID	Subgroup	N		Mean		SD		SMD HSAS	QWK		Exact Agr.	
		All	H1H2	HS	AS	HS	AS		H1H2	HSAS	H1H2	HSAS
79024	F	30555	7456	0.62	0.61	0.48	0.49	0.02	0.86	0.88	94%	94%
79024	M	30821	7593	0.59	0.58	0.49	0.49	0.02	0.88	0.89	94%	95%
79024	B	7842	1925	0.51	0.50	0.50	0.50	0.02	0.88	0.90	94%	95%
79024	H	29246	7199	0.56	0.55	0.50	0.50	0.02	0.88	0.90	94%	95%
79024	W	17707	4344	0.69	0.68	0.46	0.47	0.02	0.84	0.86	93%	94%
78742	F	31425	7817	0.57	0.58	0.50	0.49	-0.03	0.87	0.88	94%	94%
78742	M	31675	7911	0.51	0.52	0.50	0.50	-0.03	0.87	0.88	93%	94%
78742	B	8207	2024	0.41	0.42	0.49	0.49	-0.03	0.89	0.90	95%	95%
78742	H	30745	7780	0.49	0.51	0.50	0.50	-0.03	0.87	0.88	93%	94%
78742	W	17638	4346	0.62	0.64	0.49	0.48	-0.04	0.85	0.86	93%	94%
80822	F	34788	8747	0.73	0.73	0.44	0.44	0.00	0.74	0.79	90%	92%
80822	M	35130	8743	0.67	0.67	0.47	0.47	0.01	0.77	0.81	90%	92%
80822	B	8986	2253	0.63	0.64	0.48	0.48	0.00	0.79	0.81	90%	91%
80822	H	34941	8652	0.67	0.67	0.47	0.47	0.00	0.76	0.81	89%	91%
80822	W	19170	4832	0.77	0.76	0.42	0.42	0.00	0.72	0.78	90%	92%
80104	F	34971	8739	0.54	0.55	0.50	0.50	-0.02	0.81	0.87	90%	94%
80104	M	35350	8833	0.48	0.49	0.50	0.50	-0.02	0.82	0.88	91%	94%
80104	B	8871	2177	0.42	0.43	0.49	0.49	-0.02	0.83	0.89	92%	95%
80104	H	36980	9233	0.45	0.46	0.50	0.50	-0.01	0.81	0.88	91%	94%
80104	W	18013	4525	0.62	0.63	0.49	0.48	-0.03	0.79	0.85	90%	93%
81164	F	35611	9047	0.55	0.57	0.50	0.49	-0.05	0.78	0.83	89%	92%
81164	M	35064	8845	0.48	0.50	0.50	0.50	-0.06	0.76	0.82	88%	91%
81164	B	8677	2207	0.43	0.45	0.49	0.50	-0.05	0.78	0.83	89%	92%
81164	H	36826	9329	0.46	0.49	0.50	0.50	-0.05	0.78	0.83	89%	91%

Item ID	Subgroup	N		Mean		SD		SMD HSAS	QWK		Exact Agr.	
		All	H1H2	HS	AS	HS	AS		H1H2	HSAS	H1H2	HSAS
81164	W	18624	4734	0.60	0.63	0.49	0.48	-0.07	0.73	0.80	87%	91%
79244	F	36669	9174	0.36	0.36	0.48	0.48	0.00	0.76	0.82	89%	92%
79244	M	37817	9553	0.30	0.31	0.46	0.46	-0.01	0.76	0.81	90%	92%
79244	B	9736	2423	0.29	0.28	0.45	0.45	0.00	0.76	0.81	90%	92%
79244	H	40265	10088	0.30	0.30	0.46	0.46	0.00	0.75	0.81	90%	92%
79244	W	18513	4695	0.39	0.40	0.49	0.49	-0.01	0.76	0.81	88%	91%
80399	F	44898	11186	0.58	0.59	0.49	0.49	-0.01	0.92	0.94	96%	97%
80399	M	46558	11631	0.53	0.54	0.50	0.50	-0.01	0.90	0.92	95%	96%
80399	B	12080	3036	0.51	0.51	0.50	0.50	-0.01	0.92	0.93	96%	97%
80399	H	49311	12385	0.50	0.51	0.50	0.50	-0.01	0.92	0.93	96%	97%
80399	W	22654	5590	0.64	0.65	0.48	0.48	-0.02	0.90	0.93	95%	97%
81260	F	42603	10667	0.65	0.66	0.48	0.47	-0.02	0.88	0.92	94%	96%
81260	M	44779	11036	0.48	0.49	0.50	0.50	-0.01	0.90	0.92	95%	96%
81260	B	11491	2890	0.45	0.46	0.50	0.50	-0.01	0.89	0.92	95%	96%
81260	H	47728	11928	0.50	0.51	0.50	0.50	-0.01	0.89	0.92	95%	96%
81260	W	21181	5172	0.71	0.72	0.45	0.45	-0.01	0.88	0.91	95%	96%
73863	F	31675	7793	1.43	1.44	0.66	0.62	-0.02	0.72	0.78	78%	83%
73863	M	32135	8044	1.20	1.22	0.72	0.68	-0.02	0.73	0.80	75%	81%
73863	B	8074	2052	1.16	1.19	0.72	0.68	-0.05	0.72	0.78	74%	80%
73863	H	34014	8362	1.26	1.27	0.71	0.67	-0.02	0.73	0.79	76%	81%
73863	W	16182	4026	1.42	1.43	0.65	0.62	-0.01	0.72	0.79	79%	84%
68311	F	43343	10795	1.33	1.33	0.77	0.77	-0.01	0.88	0.91	86%	89%
68311	M	45255	11291	1.08	1.08	0.82	0.82	0.00	0.88	0.90	85%	87%
68311	B	11822	2919	1.10	1.09	0.80	0.80	0.01	0.88	0.90	86%	88%
68311	H	48745	12336	1.14	1.14	0.81	0.81	0.00	0.88	0.91	85%	88%
68311	W	20878	5051	1.31	1.31	0.79	0.79	-0.01	0.88	0.91	86%	89%

Table C2. STAAR Science SCR Subgroup Performance

Item ID	Subject	Subgroup	N		Mean		SD		SMD HSAS	QWK		Exact Agr.	
			All	H1H2	HS	AS	HS	AS		H1H2	HSAS	H1H2	HSAS
70928	Biology	F	36898	9222	0.67	0.66	0.68	0.68	0.00	0.91	0.92	92%	93%
70928	Biology	M	35547	8936	0.65	0.64	0.66	0.65	0.01	0.91	0.92	92%	94%
70928	Biology	B	8965	2166	0.49	0.49	0.59	0.59	0.00	0.92	0.93	94%	95%
70928	Biology	H	39440	9821	0.55	0.55	0.62	0.61	0.01	0.90	0.92	93%	94%
70928	Biology	W	17733	4560	0.84	0.84	0.70	0.70	0.00	0.89	0.91	89%	91%
70937	Biology	F	40819	10234	1.52	1.56	0.66	0.62	-0.07	0.87	0.83	90%	89%
70937	Biology	M	42639	10608	1.46	1.51	0.67	0.63	-0.07	0.85	0.83	88%	88%
70937	Biology	B	10732	2674	1.36	1.42	0.70	0.66	-0.08	0.85	0.83	88%	87%
70937	Biology	H	44262	11072	1.43	1.48	0.70	0.66	-0.07	0.87	0.84	89%	88%
70937	Biology	W	21319	5287	1.63	1.66	0.56	0.53	-0.05	0.81	0.81	90%	90%
71344	Science	F	38569	9612	0.68	0.69	0.84	0.87	0.00	0.90	0.90	87%	87%
71344	Science	M	39535	9752	0.79	0.79	0.88	0.90	-0.01	0.91	0.91	87%	87%
71344	Science	B	9922	2487	0.52	0.52	0.77	0.80	0.00	0.89	0.89	88%	87%
71344	Science	H	39748	9938	0.63	0.63	0.82	0.85	-0.01	0.89	0.89	87%	87%
71344	Science	W	20981	5120	0.94	0.95	0.89	0.92	0.00	0.91	0.91	87%	86%
60001	Science	F	34951	8697	0.74	0.79	0.73	0.75	-0.06	0.82	0.83	83%	82%
60001	Science	M	34723	8508	0.76	0.81	0.77	0.79	-0.06	0.84	0.84	82%	82%
60001	Science	B	8308	2070	0.58	0.63	0.69	0.71	-0.06	0.82	0.83	84%	84%
60001	Science	H	36522	9029	0.65	0.69	0.70	0.72	-0.06	0.81	0.82	83%	83%
60001	Science	W	18441	4559	0.91	0.97	0.80	0.81	-0.07	0.85	0.84	82%	81%
74531	Science	F	34740	8653	1.34	1.35	0.81	0.81	-0.01	0.98	0.97	98%	97%
74531	Science	M	33752	8532	1.30	1.31	0.83	0.82	-0.01	0.97	0.97	97%	97%
74531	Science	B	8169	2043	1.13	1.15	0.85	0.85	-0.01	0.98	0.97	98%	97%
74531	Science	H	35689	9028	1.21	1.22	0.84	0.84	-0.01	0.97	0.97	97%	97%
74531	Science	W	18264	4573	1.53	1.54	0.72	0.72	-0.01	0.97	0.97	98%	98%

Table C3. STAAR Social Studies SCR Subgroup Performance

Item ID	Subject	Subgroup	N		Mean		SD		SMD HSAS	QWK		Exact Agr.	
			All	H1H2	HS	AS	HS	AS		H1H2	HSAS	H1H2	HSAS
55826	U.S. History	F	31868	7896	0.96	1.04	0.76	0.73	-0.11	0.57	0.64	60%	63%
55826	U.S. History	M	33323	8229	0.84	0.87	0.75	0.73	-0.03	0.59	0.64	63%	65%
55826	U.S. History	B	8389	2121	0.83	0.84	0.76	0.72	-0.01	0.59	0.64	63%	65%
55826	U.S. History	H	34841	8552	0.83	0.88	0.75	0.73	-0.06	0.59	0.65	62%	65%
55826	U.S. History	W	17166	4256	1.02	1.09	0.75	0.72	-0.09	0.56	0.61	59%	62%
72841	U.S. History	F	34291	8508	1.26	1.30	0.71	0.66	-0.05	0.55	0.63	61%	67%
72841	U.S. History	M	33555	8511	1.18	1.19	0.73	0.69	-0.01	0.60	0.67	64%	70%
72841	U.S. History	B	8132	2029	1.09	1.10	0.73	0.68	-0.02	0.60	0.67	63%	69%
72841	U.S. History	H	35494	8895	1.15	1.16	0.72	0.67	-0.01	0.57	0.65	61%	68%
72841	U.S. History	W	18452	4660	1.36	1.39	0.68	0.64	-0.04	0.54	0.61	65%	69%
72436	Social Studies	F	37684	9437	1.40	1.39	0.77	0.81	0.00	0.80	0.83	79%	82%
72436	Social Studies	M	37278	9461	1.42	1.41	0.77	0.80	0.01	0.82	0.83	81%	82%
72436	Social Studies	B	9011	2194	1.20	1.18	0.82	0.86	0.01	0.82	0.85	79%	81%
72436	Social Studies	H	39071	9825	1.31	1.29	0.80	0.84	0.02	0.81	0.83	78%	81%
72436	Social Studies	W	19995	5149	1.62	1.64	0.64	0.66	-0.02	0.77	0.78	83%	84%
72439	Social Studies	F	36142	9084	0.64	0.65	0.80	0.81	-0.01	0.84	0.86	82%	83%
72439	Social Studies	M	35623	8957	0.67	0.66	0.81	0.80	0.00	0.84	0.85	82%	82%
72439	Social Studies	B	8604	2173	0.49	0.49	0.74	0.74	0.00	0.83	0.85	84%	85%
72439	Social Studies	H	38031	9537	0.53	0.53	0.76	0.76	0.00	0.83	0.85	83%	84%
72439	Social Studies	W	18592	4730	0.84	0.86	0.83	0.83	-0.02	0.82	0.83	78%	78%

Table C4. STAAR ECR Subgroup Performance

Item ID	Dimension	Subgroup	N	Mean		SD		SMD HSAS	QWK HSAS	Exact Agr. HSAS	Adj. Agr. HSAS	Non-Adj Agr. HSAS
				HS	AS	HS	AS					
55391	CONVENTIONS	M	23332	1.19	1.20	1.26	1.20	-0.01	0.80	59%	35%	5%
55391	CONVENTIONS	F	23889	1.39	1.40	1.34	1.27	-0.01	0.81	58%	36%	6%
55391	CONVENTIONS	B	5916	1.04	1.07	1.20	1.14	-0.03	0.79	61%	34%	5%

Item ID	Dimension	Subgroup	N	Mean		SD		SMD HSAS	QWK HSAS	Exact Agr. HSAS	Adj. Agr. HSAS	Non-Adj Agr. HSAS
				HS	AS	HS	AS					
55391	CONVENTIONS	H	22251	1.17	1.20	1.26	1.21	-0.02	0.80	60%	35%	5%
55391	CONVENTIONS	W	13955	1.43	1.40	1.31	1.23	0.02	0.80	57%	38%	6%
55391	IDEAS	M	23332	1.50	1.49	1.42	1.36	0.00	0.82	59%	34%	7%
55391	IDEAS	F	23889	1.70	1.71	1.52	1.47	-0.01	0.83	57%	35%	8%
55391	IDEAS	B	5916	1.32	1.34	1.35	1.31	-0.02	0.83	63%	31%	6%
55391	IDEAS	H	22251	1.48	1.49	1.43	1.38	-0.01	0.82	59%	34%	7%
55391	IDEAS	W	13955	1.74	1.72	1.46	1.41	0.02	0.82	56%	36%	8%
12632	CONVENTIONS	M	26720	1.25	1.23	1.46	1.41	0.01	0.82	61%	30%	9%
12632	CONVENTIONS	F	27231	1.45	1.45	1.53	1.46	0.00	0.81	58%	31%	11%
12632	CONVENTIONS	B	6430	0.99	0.97	1.36	1.30	0.01	0.82	65%	27%	8%
12632	CONVENTIONS	H	26086	1.18	1.19	1.43	1.38	-0.01	0.81	61%	29%	9%
12632	CONVENTIONS	W	15770	1.56	1.53	1.54	1.46	0.02	0.81	57%	33%	11%
12632	IDEAS	M	26720	1.61	1.52	1.76	1.69	0.05	0.85	57%	32%	11%
12632	IDEAS	F	27231	1.78	1.74	1.82	1.75	0.02	0.84	54%	34%	13%
12632	IDEAS	B	6430	1.27	1.19	1.62	1.54	0.05	0.84	62%	29%	9%
12632	IDEAS	H	26086	1.51	1.46	1.71	1.66	0.03	0.84	57%	31%	11%
12632	IDEAS	W	15770	1.92	1.85	1.83	1.75	0.04	0.84	52%	35%	12%
12638	CONVENTIONS	M	29031	1.43	1.55	1.36	1.33	-0.09	0.79	53%	38%	9%
12638	CONVENTIONS	F	27729	1.78	1.88	1.41	1.37	-0.07	0.79	51%	40%	9%
12638	CONVENTIONS	B	7210	1.23	1.35	1.27	1.24	-0.10	0.76	53%	38%	9%
12638	CONVENTIONS	H	28363	1.41	1.53	1.34	1.31	-0.09	0.78	52%	39%	9%
12638	CONVENTIONS	W	15359	1.88	1.98	1.41	1.36	-0.07	0.79	51%	41%	9%
12638	IDEAS	M	29031	1.94	1.95	1.71	1.66	0.00	0.87	53%	39%	8%
12638	IDEAS	F	27729	2.40	2.40	1.82	1.76	0.00	0.87	50%	41%	9%
12638	IDEAS	B	7210	1.69	1.68	1.58	1.48	0.01	0.85	53%	39%	7%
12638	IDEAS	H	28363	1.94	1.95	1.69	1.63	0.00	0.85	51%	40%	9%
12638	IDEAS	W	15359	2.49	2.50	1.80	1.76	0.00	0.87	49%	42%	9%
12666	CONVENTIONS	M	32333	1.77	1.76	1.57	1.57	0.01	0.86	59%	34%	8%

Item ID	Dimension	Subgroup	N	Mean		SD		SMD HSAS	QWK HSAS	Exact Agr. HSAS	Adj. Agr. HSAS	Non-Adj Agr. HSAS
				HS	AS	HS	AS					
12666	CONVENTIONS	F	32713	2.23	2.23	1.57	1.57	0.00	0.85	57%	34%	8%
12666	CONVENTIONS	B	7782	1.58	1.57	1.55	1.53	0.01	0.86	60%	33%	7%
12666	CONVENTIONS	H	33002	1.74	1.73	1.56	1.56	0.01	0.85	58%	34%	8%
12666	CONVENTIONS	W	17906	2.38	2.37	1.52	1.53	0.01	0.84	57%	35%	8%
12666	IDEAS	M	32333	2.58	2.52	2.00	1.99	0.03	0.91	53%	39%	8%
12666	IDEAS	F	32713	3.13	3.09	2.01	2.01	0.02	0.90	51%	40%	8%
12666	IDEAS	B	7782	2.30	2.26	1.94	1.92	0.02	0.90	54%	39%	7%
12666	IDEAS	H	33002	2.54	2.48	1.98	1.96	0.03	0.90	52%	40%	8%
12666	IDEAS	W	17906	3.34	3.28	1.95	1.96	0.03	0.90	51%	41%	8%
73118	CONVENTIONS	M	34356	1.81	1.83	1.53	1.52	-0.01	0.87	59%	35%	6%
73118	CONVENTIONS	F	33844	2.33	2.37	1.48	1.47	-0.03	0.85	58%	36%	6%
73118	CONVENTIONS	B	8292	1.64	1.69	1.49	1.49	-0.03	0.86	59%	35%	6%
73118	CONVENTIONS	H	35221	1.82	1.86	1.49	1.50	-0.03	0.85	57%	36%	7%
73118	CONVENTIONS	W	18264	2.47	2.48	1.47	1.45	-0.01	0.85	58%	36%	6%
73118	IDEAS	M	34356	2.43	2.49	1.88	1.80	-0.03	0.90	55%	39%	6%
73118	IDEAS	F	33844	3.06	3.12	1.89	1.80	-0.03	0.90	53%	40%	7%
73118	IDEAS	B	8292	2.22	2.31	1.82	1.74	-0.05	0.90	55%	39%	6%
73118	IDEAS	H	35221	2.44	2.51	1.82	1.74	-0.04	0.90	54%	40%	6%
73118	IDEAS	W	18264	3.22	3.24	1.90	1.81	-0.01	0.90	53%	40%	7%
73991	CONVENTIONS	M	36352	2.17	2.15	1.60	1.58	0.01	0.89	62%	32%	6%
73991	CONVENTIONS	F	37296	2.70	2.69	1.46	1.45	0.01	0.87	62%	32%	5%
73991	CONVENTIONS	B	8738	2.03	2.02	1.58	1.55	0.01	0.88	61%	33%	6%
73991	CONVENTIONS	H	38159	2.23	2.22	1.57	1.55	0.01	0.88	60%	34%	6%
73991	CONVENTIONS	W	19951	2.80	2.76	1.44	1.44	0.03	0.87	64%	31%	5%
73991	IDEAS	M	36352	2.69	2.68	2.05	2.08	0.00	0.93	57%	36%	6%
73991	IDEAS	F	37296	3.35	3.37	1.97	2.00	-0.01	0.91	55%	39%	7%
73991	IDEAS	B	8738	2.49	2.49	2.01	2.02	0.00	0.92	58%	37%	6%
73991	IDEAS	H	38159	2.73	2.75	2.00	2.04	-0.01	0.92	56%	37%	6%

Item ID	Dimension	Subgroup	N	Mean		SD		SMD HSAS	QWK HSAS	Exact Agr. HSAS	Adj. Agr. HSAS	Non-Adj Agr. HSAS
				HS	AS	HS	AS					
73991	IDEAS	W	19951	3.49	3.48	1.96	1.99	0.01	0.91	55%	39%	6%
68583	CONVENTIONS	M	45360	1.98	2.10	1.51	1.46	-0.08	0.85	55%	39%	7%
68583	CONVENTIONS	F	43376	2.48	2.61	1.44	1.39	-0.09	0.84	56%	38%	6%
68583	CONVENTIONS	B	11299	1.82	1.95	1.47	1.42	-0.09	0.83	53%	40%	7%
68583	CONVENTIONS	H	48030	1.97	2.10	1.49	1.45	-0.09	0.84	54%	39%	7%
68583	CONVENTIONS	W	21976	2.75	2.86	1.35	1.28	-0.08	0.82	57%	37%	6%
68583	IDEAS	M	45360	2.69	2.77	1.97	2.01	-0.04	0.91	55%	39%	7%
68583	IDEAS	F	43376	3.35	3.44	1.94	1.94	-0.05	0.90	53%	40%	7%
68583	IDEAS	B	11299	2.43	2.50	1.89	1.93	-0.04	0.91	55%	39%	6%
68583	IDEAS	H	48030	2.67	2.77	1.94	1.98	-0.05	0.90	54%	39%	7%
68583	IDEAS	W	21976	3.73	3.80	1.84	1.84	-0.04	0.89	53%	40%	6%
68776	CONVENTIONS	M	40116	2.06	2.16	1.58	1.46	-0.06	0.84	54%	37%	8%
68776	CONVENTIONS	F	40274	2.60	2.66	1.49	1.35	-0.05	0.82	55%	37%	8%
68776	CONVENTIONS	B	10168	1.95	2.05	1.55	1.42	-0.07	0.82	52%	39%	9%
68776	CONVENTIONS	H	43004	2.08	2.18	1.57	1.45	-0.07	0.83	54%	38%	8%
68776	CONVENTIONS	W	20547	2.83	2.85	1.39	1.26	-0.01	0.80	56%	36%	8%
68776	IDEAS	M	40116	2.63	2.63	2.06	2.06	0.00	0.91	53%	39%	8%
68776	IDEAS	F	40274	3.36	3.42	2.03	2.01	-0.03	0.90	51%	40%	8%
68776	IDEAS	B	10168	2.45	2.47	1.98	1.98	-0.01	0.90	52%	40%	8%
68776	IDEAS	H	43004	2.68	2.73	2.05	2.06	-0.02	0.91	52%	39%	8%
68776	IDEAS	W	20547	3.62	3.62	1.96	1.95	0.00	0.90	51%	41%	8%