Texas STAAR RLA Spring 2024 Administration: Automated Scoring Methods and Results

Executive Summary

Overall, the results suggest that the hybrid scoring design is providing accurate, reliable, and fair scoring. All items scored in Spring 2024 met our full set of performance criteria on the full random sample.

Routing for both low confidence and condition code routing are performing adequately. The low confidence routing performances indicate that the engine is not performing well on these responses, which suggests that the confidence model and threshold is identifying responses that are difficult to score and should be routed for human scoring. The condition code routing agreements indicate that responses scored with the Out of Vocabulary condition code show very high agreements with the human raters. The other two condition codes performed adequately but will continue to be refined to improve agreements with the human raters.

Areas of future consideration include research into further refining the overall hybrid scoring design. This includes ensuring that hand-scores are returned quickly enough to reprogram the engine earlier in the test window. It also includes examining the impact of not using the original model for routing low confidence or condition code responses, and instead reserving that routing only for the final reprogrammed model. In order to ensure that 25% of responses are routed under this approach we can examine whether to increase the threshold for low confidence routing or increase the percentage of responses in the random percent routed sample. We will also examine changing the Unusual Score condition code to allow for these responses to be routed to the typical human rater pool, rather than the expert rater pool. The Out of Vocabulary condition code could potentially be considered for non-routing.

Table of Contents

Introduction

Cambium Assessment, Inc. (CAI) and Pearson under the direction of TEA assessment staff conducted hybrid automated/human scoring of STAAR items administered in English and of TELPAS items in grades 4 and above for all constructed response items during the 2023-2024 school year. This technical report focuses on the constructed response items included in the **Spring STAAR Reading Language Arts (RLA) Grades 3-8 and EOC** assessments. The STAAR RLA program has 24 constructed response items: 16 short constructed response (SCR) items and 8 extended constructed response (ECR, or essay) items. Separate reports discuss STAAR Science and Social Science assessments, EOC assessments administered in December 2023 and June 2024, and TELPAS.

The purpose of this technical report is to document CAI's procedures and to examine the performance of CAI's automated scoring engine, ASE, relative to human scoring when evaluating models in the hybrid scoring process. The hybrid scoring method was based upon a study conducted on Spring 2023 data; the technical report for this is entitled "The State of Texas Assessments of Academic Readiness (STAAR) Hybrid Scoring Study Methods and Results: Spring 2023 Items" and is available on the Texas website.

The hybrid scoring method has multiple steps. First, ASE models are programmed on data from the most recent test administration for that item; these data could come from a Stand Alone Field Test (SAFT) administration, an embedded field test item or an operational item in an operational administration. Once deployed for operational scoring, all responses receive scores from ASE. Approximately 25% of responses are routed for independent human scoring using three routing rationales: random, condition code, and low confidence. When routed for human scoring, the human score is the final reported score.

During test administration, the performance of ASE and of human scoring on each item is monitored daily. All item models are reprogrammed using the operational responses and scores in the randomly sample of responses. Models are reprogrammed on the operational data to ensure that scores produced by the engine reflect how students are writing and how programmed human raters are scoring responses in that administration. This approach was recommended by the TEA technical advisory committee based upon the Spring 2023 report. Once reprogrammed on the operational responses and scores, all responses are rescored. Any new condition codes or low confidence responses produced by the reprogrammed model are routed for human scoring. Responses receiving a human score, either as routed by the original or reprogrammed model, retain that human score as the final score of record.

In Spring 2024, a total of 9,690,388 STAAR RLA responses were scored using the hybrid scoring approach. 72.2% of responses received scores from ASE alone, and 28.2% were routed for human scoring and received those scores as final reported scores.

This technical report focuses primarily on the operational reprogrammed models and scores. We begin by describing the methodology of ASE programming, hybrid design, and how the scoring performance is evaluated. Then, we present the results. We end with recommendations, particularly around the implementation of the hybrid design.

Methods

We briefly describe the constructed response items, student-level data sources, hand-scoring procedures, automated scoring methods, and metrics used to evaluate the automated scoring engine.

Items

The Spring 2024 STAAR 3-8 and EOC RLA assessments consisted of a mix of SCR and ECR items. Each grade included 2 SCRs and 1 ECR item. Across all grades, there were a total of 24 RLA items in STAAR. The SCR items were of two types: One-point SCR items asked students to rewrite one or more sentences for clarity and correctness; two-point SCR items asked students to respond to a reading comprehension prompt after reading a passage. The ECR items are essay items, scored in two dimensions: Ideas and Conventions. The Ideas dimension rubric ranged from 0 to 3, and the Conventions rubric ranged from 0 to 2. Notably, students could earn scores of 0 in the Ideas and Conventions rubrics even when providing a valid response. For instance, students could have a controlling idea, but lack an introduction or conclusion and have little or no idea expression or organizational structure and earn a score of 0. Additionally, students receiving 0 in Ideas also received a 0 score in Conventions, according to the rubric. Finally, scores in Ideas and Conventions are not reported on the rubric scale; rather, they are reported as the sum of two rater scores or twice an expert read.

All items administered in RLA assessments are presented in Table 1, along with information on the item type, maximum rubric score, and most recent administration for that item prior to Spring 2024. Recall that items, once administered operationally, are typically released in the STAAR Spring assessment program. The recent administration type reflects the data sources used for programming the engine prior to the start of the Spring 2024 administration.

Table 1. SCR and ECR items administered as a part of the Spring 2024 STAAR 3-8 RLA and EOC assessment

Grade	Item ID	Item Type	Dim.	Max Score	Most Recent Administration
3	114749	SCR	Overall	1	EFT 2023
3	83640	SCR	Overall	2	EFT 2023
3	12624	ECR	Conv. Ideas	4 6	SAFT 2022
4	114768	SCR	Overall	1	EFT 2023
4	91650	SCR	Overall	2	EFT 2023
4	12628	ECR	Conv. Ideas	4 6	SAFT 2022
5	114786	SCR	Overall	1	SAFT 2022
5	84308	SCR	Overall	2	EFT 2023
5	12647	ECR	Conv. Ideas	4 6	SAFT 2022

Grade	Item ID	Item Type	Dim.	Max Score	Most Recent Administration
6	114807	SCR	Overall	1	EFT 2023
6	2224	SCR	Overall	2	SAFT 2022
6	12674	ECR	Conv. Ideas	4 6	SAFT 2022
7	114822	SCR	Overall	1	EFT 2023
7	90459	SCR	Overall	2	EFT 2023
7	61507	ECR	Conv. Ideas	4 6	SAFT 2022
8	114840	SCR	Overall	1	EFT 2023
8	89173	SCR	Overall	2	EFT 2023
8	73974	ECR	Conv. Ideas	4 6	SAFT 2022
9	113231	SCR	Overall	1	EFT 2023
9	90632	SCR	Overall	2	EFT 2023
9	68219	ECR	Conv. Ideas	4 6	SAFT 2022
10	113258	SCR	Overall	1	EFT 2023
10	89405	SCR	Overall	2	EFT 2023
10	69030	ECR	Conv. Ideas	4 6	SAFT 2022

Note: EFT refers to an embedded field test item; SAFT refers to the standard alone field test. The year (e.g. 2022, 2023) refers to the year in which the spring administration occurred.

Data

Data for the hybrid scoring model comes from two key sources. The first source is the data used to program the models initially; as noted in the Items section, these data came from the EFT or SAFT administrations from prior years. The second source is the data from the Spring 2024 operational administration. This administration occurred between 4/7/2024 and 4/19/2024. The in-window reprogramming was based on a subset of these responses—the 10% random sample of the first wave of test-taker responses for which hand scores were available. Approximately 15% of this sample was held out to determine model performance. See the Model Programming section for more details regarding in-window reprogramming.

Hybrid Scoring Approach

The hybrid scoring approach results in responses ultimately receiving a score from ASE or programmed human raters. All responses receive scores from ASE. Approximately 25% of responses are routed for human scoring; when routed for human scoring, the human score is considered the final score. Responses routed for human scoring do not include the engine score to ensure independence between the human and engine scoring. Additionally, responses routed for human scoring also receive a percentage of second reads to examine how well the humans are agreeing with one another.

The hybrid scoring design has multiple steps. These are described in order below.

1. Responses which receive algorithmic condition codes defined for each item and type (e.g., responses to ECR items with fewer than 9 words are assigned a condition code of NOT_ENOUGH_DATA). These responses receive a score of 0 and are not routed for human scoring. The condition codes and thresholds are defined by TEA and are based on both empirical evidence of engine performance and a content-based judgement about the responses that do not meet the minimum rubric criteria.
2. Approximately 10% of responses are routed for a random scoring verification check to monitor engine and human scoring performance.
3. Responses assigned condition codes by ASE that are indicative of unusual response patterns or scores are routed for humans to provide a final score. These responses are routed for expert scoring.
4. Responses that receive low confidence percentile scores from the engine (less than the 10th percentile) are routed for humans to provide the final score. These low confidence responses reflect scores that ASE has deemed as having low likelihood of matching an expert human score.

Because the hybrid design—particularly when routing condition codes and low confidence responses—was influenced by two models, we expect the overall routing percentages for condition codes and low confidence to be higher across the two models than for any individual model. For instance, the low confidence percentile threshold of 10% will flag approximately 10% of responses for the original model and 10% of responses for the reprogrammed model. Because the responses are rescored using the reprogrammed model, we also expect some overlap between the two model results, meaning that responses could be flagged as low confidence by both models. We can also expect that a response may be flagged as low confidence under the original model but not the reprogrammed model. Or a response may be flagged as low confidence under the reprogrammed model but not the original model. Finally, a response may not be flagged by either model as low confidence. Regardless, any response routed for human scoring will have the score of the reprogrammed model and the human score, with the human score serving as the score of record. This same logic exists for condition codes that are routed.

The non-routed ASE condition codes and the random routing are not affected by the reprogrammed model and rescore because these are deterministic processes that are not impacted by model reprogramming.

ASE Description

ASE 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.

For ECR items, two independent models were programmed to score each dimension. Thus, two models were programmed to score Ideas, and another two models were programmed to score Conventions. All models were programmed to predict single rater scores as the dependent variable. More specifically, model 1 (M1) was programmed to predict human rater 1 scores (H1), and model 2 (M2) was programmed to predict human rater 2 scores (H2). For SCR items, two independent models (M1 and M2) were programmed to score each item; each model was programmed on the final resolved score rather than the two rater scores in order to ensure each model was programmed on the best available score. For each item and dimension, the two models were combined via *ensembling* to generate the final score.

ASE 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. These are described in detail in their respective sections.

Combining Models

In ASE, 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. It is particularly effective when the models are different from each other.

For SCR items, the ensembling mechanism is logistic regression, using the output logits or probabilities from M1 and M2. In the case of ECR items, M1 outputs are combined with M2 outputs to produce a final score that reflects the summed score, essentially simulating the human rater scoring process. Because the final dimension score is a sum of H1 and H2, the output probabilities of M1 and M2 were combined to produce a probability distribution on the same scale as the final dimension score.[1] The max probability was taken as the final dimension score. Consistent with human rater scoring procedures, final dimension scores were summed to create the final item score.

Condition Codes

ASE produces condition codes as part of its scoring process. Condition codes are used to identify responses that do not meet the minimal rubric requirements or that should be routed for human scoring. The choice of condition codes, their thresholds, and routing decisions were decided upon with TEA using the Spring 2023 data.

ASE produced nine condition codes. Table 2 lists these condition codes with a description, which item type for which the condition code is used and whether the condition code is routed for human scoring. Any response receiving a non-routed condition code is assigned a score of 0 overall and in each dimension.

As noted in the table, responses receiving condition codes NO_RESPONSE, COMMON_REFUSAL, NON_SCORABLE_LANGUAGE, NOT_ENOUGH_DATA, DUPLICATE_TEXT, and PROMPT_COPY_MATCH were not routed for human scoring. Responses receiving the OUT_OF_VOCAB, NON_SPECIFIC, and UNUSUAL_SCORES condition code were routed for human scoring.

[1] This summation occurs on the model probabilities, whereby the probability of the summed score is the sum of the products of the model probabilities for all possible sums for 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 score associated with the highest probability.

Table 2. Condition codes employed in the Spring 2024 STAAR RLA assessment

ASE Condition Code	Description	Applies to	Routed for Human Scoring
NO_RESPONSE	No non-blank characters are detected in the response.	SCR ECR	No
COMMON_ REFUSAL	Response only contains words associated with a refusal such as 'I don't know' or contains only non-alphanumeric characters.	SCR ECR	No
NON_ SCORABLE_ LANGUAGE	Response is longer than 30 characters and is written primarily in Spanish.	SCR ECR	No
NOT_ ENOUGH_ DATA	Student response is less than the minimum number of words configured in the rubric.	1-pt. SCR ECR	No
DUPLICATE_ TEXT	Student response consists primarily of text copied over and over.	SCR ECR	No
PROMPT_ COPY_ MATCH	Student response is primarily copied from the passage or item prompt. Percentage of characters in the response that appear in the passage.	ECR	No
OUT_OF_ VOCAB	The ratio of the sum of the lengths of words in a response that are in the engine programming sample over the sum of length of all words in the response	SCR ECR	Yes
UNUSUAL_ SCORES	Identifies responses with ASE scores that are unusual in some way (i.e., non-adjacent or very short but receiving greater than the minimum rubric score.	ECR	Yes
NONSPECIFIC	Essay scoring engine predicts the assignment of a condition code using a statistical procedure (not threshold).	SCR ECR	Yes

Confidence

ASE produces confidence values as part of its scoring process. The confidence value reflects the degree to which ASE is confident in the score it has predicted. A high confidence value indicates that the engine is confident that its predicted score matches the score of a final human score; a low confidence value indicates that the engine is less confident that its predicted score matches the score of a final human score. The confidence values are reported as percentiles.

The confidence model is programmed (using probit regression) to predict whether the engine score matches the final human score on a 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 to have a mean score of 0 and standard deviation of 1, and then summed to provide an overall item confidence score.

ASE Model Programming

For Texas assessment programs, ASE models were programmed in two phases. In the first phase, models were programmed on the EFT or SAFT data from a prior test administration. In the second phase, models were reprogrammed on operational data from the in-window testing administration. As noted in the introduction, the models programmed on the operational data produced the final scores. This two-phase process was used to ensure that models were available for scoring at the start of the test administration and that the highest quality models were used to score Texas student responses. While the model programmed on the EFT or SAFT data can perform well compared to programmed human raters, the model programmed on the operational data is typically programmed on more data that reflects the actual responses of Texas students during live testing. For this reason, the models reprogrammed on the operational data are prioritized, even when the EFT/SAFT models perform well.

For both phases, CAI programs models for each item and dimension. Data are divided into programming and held-out validation sets, with 70% of responses used to program the engine models, 15% to program the ensembler[2], and 15% used to evaluate the engine performance. The held-out validation data were also used to program the confidence models and to build the confidence percentiles. 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 ASE condition codes.

Hand-scoring Procedures

The technical digest describes procedures around programming human raters for the STAAR program. This document focuses on the second read and resolution rules implemented in hand-scoring, as these serve as the basis for monitoring automated scoring performance. Recall that engine scores are not routed along with responses; the hand-scoring procedures operate independently of the engine scoring. This approach supports the ability to compare ASE human performance on an independent sample and to use these data for engine reprogramming.

When responses are routed for hand-scoring, the scoring process varies by item type. Table 3 presents these data.

Table 3. Hand-scoring reliability reads and resolution rules

Item Type	% Reliability Read	Resolution	Final Score
1 point SCR	25%	Any non-exact score is resolved by an expert reader	Reader 1 or Expert Reader
2 point SCR	25%	Any non-exact score is resolved by an expert reader	Reader 1 or Expert Reader

[2] Note that the ECR ensembler does not require estimation of parameters because it sums the outputs from the two models. Even so, we retain the data structure and methods for simplicity across items.

ECR	100%	Any non-adjacent score is resolved by an expert rater, within dimension	Sum of Reader 1 and Reader 2 or double Expert Reader

In addition to Initial, Reliability, and Expert scores, a small percentage of responses received backread scores assigned by the supervisor, typically as quality checks of human raters. These scores also serve as the score of record, if they exist.

The condition codes used by the human raters appears in Table 4.

Table 4. List of human rater condition codes

Condition Code	Definition
B	Blank
C	Lacks any original writing
D	Insufficient response
F	Written in a language other than tested language
I	Indecipherable
P	Does not write in prose
R	Refuses to write
T	Off topic

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 or QWK using Fleiss-Cohen weights) and a distributional measure (Standardized Mean Difference or SMD using pooled standard deviation). Each of these are described in greater detail below.

CAI used the following thresholds to identify poorly performing items:

- Engine-Final, resolved score exact agreement lower than 5.25% of human-human exact agreement (PARCC, 2015)
- Engine-Final, resolved QWK lower than .10 of human-human QWK (Williamson et al., 2012)
- Engine-Final, resolved SMD magnitude greater than .15 (Williamson et al., 2012).

For the STAAR ECR summed scores, there is no comparable H1H2 agreement and so only two measures are used:

- Engine-Final, resolved QWK less than .7 (Williamson et al., 2012)
- Engine-Final, resolved SMD magnitude greater than .15 (Williamson et al., 2012).

For STAAR ECR items, we focus on the summed score evaluation when evaluating overall performance. However, we also examine performance of each model on the rubric score for each dimension as well.

The application of the metrics was conducted on the sample of response in which both ASE and human-assigned condition codes were removed. This approach was taken because the core focus is on the ability of the engine to reproduce rubric scores.

Exact Agreement

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. Typically, human-machine (HSAS) exact agreement should be no less than 5.25% the human-human (H1H2) exact agreement rate (PARCC, 2015).

Quadratic Weighted Kappa

Also referred to as Cohen's kappa, or a kappa value, QWK provides a measure of agreement where a value of 1 represents perfect agreement and a value of 0 indicates random chance. QWK uses the Fleiss-Cohen weights. As indicated by its name, QWK weights disagreements as the square of the difference in points, relative to the score range. Hence, QWK penalizes large disagreements much more than small disagreements. Typically, HSAS QWK should be no less than .10 H1H2 QWK (Williamson et al., 2012).

Standardized Mean Difference

SMD examines whether two rater groups are scoring differently from one another without having to know the scale of a particular item. To calculate 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 using the pooled standard deviation. 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 HSAS SMDs to differ by no more than a magnitude of 0.15 (Williamson et al., 2012).

Evaluation Approach

The performance of automated scoring in the hybrid scoring model was evaluated on several different samples. All evaluations used the evaluation metrics on the defined sample.

First, the hybrid scoring model percentages routed for human scoring are evaluated relative to expected performance. Second, the performance of the reprogrammed model on the held-out validation data is evaluated. Third, the performance of the reprogrammed model on the full random routed sample is evaluated. Fourth, the performance of the engine scores on the condition code-routed responses is evaluated. Fifth, the performance of the engine scores on the low confidence routed responses is evaluated. Finally, the overall score distribution on the full sample—whether AS or human scored—is presented and compared to the random sample. This analysis illustrates the degree to which the random sample represents the full sample set of scores.

During testing, four different monitoring reports were generated and reviewed daily. These were provided daily to TEA and discussed at regular intervals throughout and at the end of the testing window.

1. **Routing Report**: A high level overview report detailing the percentage routed for the three different routing codes (random, condition code, and low confidence). This includes counts and percentage of the total number of students tested for each item. This also includes the number and percentage of hand-scores returned.
2. **Performance Report**: This Performance Report provides details regarding model performance on the random sample (of approximately 10% of responses) that were routed for human scoring, excluding essays that were flagged by one of the condition codes.
3. **Routing Code Analysis Report**: This report provides various measures of agreements for three routing conditions used in the STAAR assessment.
4. **Performance Report for essays routed for Low confidence**: This Performance Report provides details regarding model performance on the responses that were flagged as being low confidence, and routed for human scoring, excluding responses that were flagged by one of the condition codes. We expect these responses to have lower agreement with human raters precisely because lower confidence values mean that the confidence model predicts the engine scores to be less likely to match the human scores compared to higher confidence values.

Results

Results are organized broadly around the samples collected throughout the automated scoring process. Specifically, these include the held-out validation sample (i.e., the held-out data from in-window reprogramming), the three routed samples (Random Percent, Condition Codes that require human review, and Low Confidence), and all scored responses among all test-takers. The Results section begins with the number and percentage of responses that were routed. We describe the results on the held-out validation sample, followed by the random percent sample, since these bear most directly on the performance of ASE. Then we consider routed condition codes, low-confidence scores, and final scores among all test-takers.

Routing Percentages

In this section, we present the number and percentage of responses routed for hand-scoring under the three routing conditions for SCR items and for ECR items. Recall that, for condition codes and low confidence, responses could be routed using output from the original or reprogrammed model.

Table 5 presents the number and percentage of responses routed for human scoring for SCR items for all responses. Across SCR items, between 9.5% and 10% of all responses were randomly routed for human scoring. Between 0.8% and 9.1% of responses were routed due to condition codes assigned by either ASE model. Between 12.2% and 21.1% of response were routed due to low confidence percentile values from either model being below the 10th percentile. Between 24.1% and 36.8% of SCR responses were routed for human scoring.

Table 5. Number and percentage of responses routed for human scoring across all responses for SCR items

Grade	Item ID	Max Score	N Total	Random Sample		Condition Code		Low Confidence		All Routed	
				N	%	N	%	N	%	N	%
3	114749	1	356,885	34,103	9.6%	9,285	2.6%	43,402	12.2%	86,790	24.3%
3	83640	2	358,763	35,504	9.9%	32,716	9.1%	63,833	17.8%	132,053	36.8%
4	114768	1	366,410	35,570	9.7%	6,832	1.9%	54,393	14.8%	96,795	26.4%
4	91650	2	367,910	36,911	10.0%	5,442	1.5%	63,919	17.4%	106,272	28.9%
5	114786	1	374,249	36,979	9.9%	5,419	1.4%	68,051	18.2%	110,449	29.5%
5	84308	2	374,952	37,489	10.0%	4,452	1.2%	71,441	19.1%	113,382	30.2%
6	114807	1	392,519	38,601	9.8%	3,890	1.0%	63,103	16.1%	105,594	26.9%
6	2224	2	393,258	39,129	9.9%	5,074	1.3%	83,019	21.1%	127,222	32.4%
7	114822	1	396,050	38,686	9.8%	3,684	0.9%	59,890	15.1%	102,260	25.8%
7	90459	2	396,721	39,403	9.9%	2,815	0.7%	69,562	17.5%	111,780	28.2%
8	114840	1	401,068	39,405	9.8%	3,028	0.8%	63,580	15.9%	106,013	26.4%
8	89173	2	401,408	39,744	9.9%	3,510	0.9%	62,126	15.5%	105,380	26.3%
9	113231	1	482,703	46,035	9.5%	4,475	0.9%	75,321	15.6%	125,831	26.1%
9	90632	2	484,614	47,717	9.8%	4,755	1.0%	80,524	16.6%	132,996	27.4%
10	113258	1	459,933	44,347	9.6%	3,554	0.8%	62,885	13.7%	110,786	24.1%
10	89405	2	460,370	45,357	9.9%	4,577	1.0%	71,885	15.6%	121,819	26.5%
		Total	6,467,813	634,980	9.8%	103,508	1.6%	1,056,934	16.3%	1,795,422	27.8%

Table 6 presents the number and percentage of responses routed for human scoring for ECR items for all responses. Across ECRs, between 8.5% and 9.3% of all responses were routed for human scoring randomly. Between 0.6% and 10.3% of responses were routed due to condition codes assigned by either ASE model. Between 12.0% and 24.3% of response were routed due to low confidence percentile values from either model being below the 10th percentile. Between 23.6% and 37.5% of ECR responses were routed for human scoring.

Table 6. Number and percentage of responses routed for human scoring across all responses for ECR items

Grade	Item ID	N Total	Random Sample		Condition Code		Low Confidence		All Routed	
			N	%	N	%	N	%	N	%
3	12624	357,552	30,354	8.5%	22,374	6.3%	53,564	15.0%	106,292	29.7%
4	12628	366,764	32,738	8.9%	24,404	6.7%	45,565	12.4%	102,707	28.0%
5	12647	374,151	34,927	9.3%	38,715	10.3%	66,749	17.8%	140,391	37.5%
6	12674	392,073	35,972	9.2%	7,589	1.9%	69,199	17.6%	112,760	28.8%
7	61507	395,432	36,876	9.3%	2,294	0.6%	96,136	24.3%	135,306	34.2%
8	73974	399,453	36,462	9.1%	2,906	0.7%	75,873	19.0%	115,241	28.8%
9	68219	478,949	42,377	8.8%	9,309	1.9%	69,710	14.6%	121,396	25.3%

Grade	Item ID	N Total	Random Sample		Condition Code		Low Confidence		All Routed	
			N	%	N	%	N	%	N	%
10	69030	458,201	41,820	9.1%	11,372	2.5%	54,817	12.0%	108,009	23.6%
	Total	3,222,575	291,526	9.0%	118,963	3.7%	531,613	16.5%	942,102	29.2%

Non-Routed Condition Codes

The percentage of responses receiving condition codes from ASE that were not eligible for human routing appears in Table 7 for SCR items and Table 8 for ECR items. Recall that these responses received 0s overall and in each domain for ECRs. For SCR items, the percentages ranged from 0.0% to 3.4%.

Table 7. Percentage of responses receiving non-routed condition codes for all SCR items

Grade	Item ID	Max Score	N Total	Percent Non-Routed Condition Codes					
				No Response	Common Refusal	Non-Scorable Language	Not Enough Data	Duplicate Text	Prompt Copy Match
3	114749	1	356,885	0.1%	0.4%	0.0%	3.4%	0.0%	
3	83640	2	358,763	0.3%	0.4%	0.0%		0.0%	
4	114768	1	366,410	0.1%	0.3%	0.0%	2.2%	0.0%	
4	91650	2	367,910	0.1%	0.2%	0.0%		0.0%	
5	114786	1	374,249	0.1%	0.3%	0.0%	1.1%	0.0%	
5	84308	2	374,952	0.1%	0.2%	0.0%		0.0%	
6	114807	1	392,519	0.1%	0.5%	0.1%	1.3%	0.0%	
6	2224	2	393,258	0.1%	0.4%	0.2%		0.0%	
7	114822	1	396,050	0.1%	0.6%	0.1%	1.2%	0.0%	
7	90459	2	396,721	0.1%	0.5%	0.1%		0.0%	
8	114840	1	401,068	0.1%	0.7%	0.1%	1.0%	0.0%	
8	89173	2	401,408	0.1%	0.7%	0.1%		0.0%	
9	113231	1	482,703	0.1%	1.6%	0.0%	2.3%	0.0%	
9	90632	2	484,614	0.1%	1.2%	0.1%		0.0%	
10	113258	1	459,933	0.1%	1.4%	0.0%	1.8%	0.0%	
10	89405	2	460,370	0.2%	1.6%	0.0%		0.0%	
		Total	6,467,813	0.1%	0.7%	0.1%	0.9%	0.0%	

Entries are blank when condition codes are not applied to the given item.

For ECR items, the percentages of non-routed condition codes assigned were higher. Much of the increase was due to the NOT ENOUGH DATA condition code (when responses are eight or fewer words) or the PROMPT COPY MATCH condition code (when 80% of the response exactly matches the passage, prompt, or directions).

Table 8. Percentage of responses receiving non-routed condition codes for all ECR items

Grade	Item ID	N Total	Percent Non-Routed Condition Codes					
			No Response	Common Refusal	Non-Scorable Language	Not Enough Data	Duplicate Text	Prompt Copy Match
3	12624	357,552	0.2%	0.4%	0.1%	9.6%	0.0%	5.4%
4	12628	366,764	0.2%	0.4%	0.0%	4.6%	0.0%	6.1%
5	12647	374,151	0.1%	0.4%	0.0%	3.4%	0.0%	3.4%
6	12674	392,073	0.2%	0.7%	0.2%	4.3%	0.0%	2.9%
7	61507	395,432	0.2%	0.8%	0.1%	3.0%	0.0%	3.1%
8	73974	399,453	0.3%	1.2%	0.1%	3.7%	0.0%	3.4%
9	68219	478,949	0.4%	2.6%	0.1%	5.7%	0.0%	3.6%
10	69030	458,201	0.2%	1.8%	0.0%	3.5%	0.0%	3.6%
	Total	3,222,575	0.2%	1.1%	0.1%	4.7%	0.0%	3.9%

Entries are blank when condition codes are not applied to the given item.

Operational Held-Out Validation Sample

The operational held-out validation sample refers to the held-out data from in-window reprogramming. All item models were reprogrammed a few days after the test administration closed on a subset of data from the random routed sample for which hand-scores were available.

ASE Programming and Validation Sample

Table 9 and Table 10 show the number of responses used to program ASE and validate ASE performance as well as the total size of the random sample for the entire administration. The total number of responses varies by item because the hand-scoring rate varies by item, due to when tests are administered to students, hand-scoring programming and resourcing, and the complexity of the item. Note that the total number of responses used in ASE programming and validation represents an average of 28.8% of random routed responses for SCR items (ranging from 14.5% to 50.4%) and an average of 17.4% of random routed responses for ECR items (ranging from 6.1% to 25.5%).

Table 9. Number and percentage of responses used to program and validate ASE performance on SCR Items

Grade	Item ID	Max Score	Random Sample	Reprogramming Sample	
			N	N	%
3	114749	1	33,514	9,179	27.4%
3	83640	2	34,273	6,858	20.0%
4	114768	1	35,110	9,401	26.8%
4	91650	2	36,482	6,869	18.8%
5	114786	1	36,724	8,543	23.3%
5	84308	2	37,194	5,394	14.5%
6	114807	1	38,393	11,862	30.9%

Grade	Item ID	Max Score	Random Sample	Reprogramming Sample	
			N	N	%
6	2224	2	38,821	9,634	24.8%
7	114822	1	38,537	11,885	30.8%
7	90459	2	39,190	10,308	26.3%
8	114840	1	39,303	13,044	33.2%
8	89173	2	39,478	12,895	32.7%
9	113231	1	45,861	23,129	50.4%
9	90632	2	47,355	14,577	30.8%
10	113258	1	44,213	15,078	34.1%
10	89405	2	45,037	12,916	28.7%
		Total	629,485	181,572	28.8%

Note: Reprogramming Sample refers to the in-window random sample collected and used for the reprogramming of final models. This sample does not include condition codes.

Table 10. Number and percentage of responses used to program and validate ASE Performance on ECR Items

Grade	Item ID	Random Sample	Reprogramming Sample	
		N	N	%
3	12624	29,625	5,720	19.3%
4	12628	31,955	8,144	25.5%
5	12647	34,128	7,978	23.4%
6	12674	35,671	5,569	15.6%
7	61507	36,722	5,237	14.3%
8	73974	36,407	7,045	19.4%
9	68219	42,086	8,001	19.0%
10	69030	41,233	2,515	6.1%
	Total	287,827	50,209	17.4%

Note: Reprogramming Sample refers to the in-window random sample collected and used for the reprogramming of final models. This sample does not include condition codes.

ASE Performance

ASE performance on the operational held-out validation sample compares the final score and engine agreement (HSAS) to the agreement of the two humans when second reads are conducted. The human-human (H1H2) agreements are from the full random sample for SCR items and are from the held-out validation sample for the ECR items.[3] The HSAS agreements are from the held-

[3] The data provided for engine reprogramming consisted primarily of responses in which the two human raters agreed. This resulted in falsely high H1H2 agreement estimates. Thus, we use the full sample for H1H2 evaluation in this report. During live testing, we used estimates provided from Pearson on hand-scoring data that had not yet been made available for engine programming. These estimates indicated that the engine was performing well on each item.

out validation sample. SCR item results are presented first, followed by the ECR item results. For score point distribution performance, please see Appendix A.

Table 11 presents the H1H2 and HSAS exact agreements and QWK values for SCR items. Across nearly all items and measures, values were similar between H1H2 and HSAS. One item in Grade 10, however, did not meet the exact agreement evaluation criteria but was close to the threshold.

Table 11. Performance of ASE with respect to Exact Agreement and QWK on SCR items in the held-out validation sample

Grade	Item ID	N HSAS	N H1H2	Max Score	Exact Agreement			QWK		
					H1H2	HSAS	diff	H1H2	HSAS	diff
3	114749	1,363	6,138	1	96.3%	96.3%	0.0%	0.92	0.93	0.01
3	83640	1,008	7,056	2	77.9%	82.7%	4.8%	0.79	0.84	0.05
4	114768	1,401	6,435	1	96.1%	96.7%	0.6%	0.91	0.93	0.02
4	91650	1,024	10,315	2	67.9%	72.6%	4.7%	0.68	0.76	0.08
5	114786	1,280	6,334	1	91.2%	93.9%	2.7%	0.82	0.88	0.06
5	84308	800	10,608	2	71.5%	78.5%	7.0%	0.73	0.8	0.07
6	114807	1,774	7,438	1	91.6%	93.6%	2.0%	0.82	0.86	0.04
6	2224	1,439	10,091	2	75.6%	80.7%	5.1%	0.78	0.84	0.06
7	114822	1,782	6,868	1	96.8%	98.1%	1.3%	0.93	0.96	0.03
7	90459	1,541	9,954	2	73.8%	79.5%	5.7%	0.75	0.83	0.08
8	114840	1,955	6,971	1	89.7%	93.2%	3.5%	0.78	0.85	0.07
8	89173	1,933	10,856	2	76.0%	82.9%	6.9%	0.77	0.85	0.08
9	113231	3,462	8,992	1	97.2%	98.6%	1.4%	0.94	0.97	0.03
9	90632	2,161	10,334	2	79.3%	85.1%	5.8%	0.80	0.85	0.05
10	113258	2,260	8,918	1	92.0%	94.9%	2.9%	0.84	0.89	0.05
10	89405	1,919	11,878	2	84.1%	77.0%	-7.1%	0.85	0.79	-0.06
				Avg.	84.8%	87.8%	3.0%	0.82	0.86	0.04

Note: For SCR items, target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

Table 12 presents the HS and AS means and standard deviations, as well as the SMD values for SCR items. For all items, the SMD values were within performance thresholds.

Table 12. Performance of ASE with respect to SMD on SCR items in the held-out validation sample

Grade	Item ID	N HSAS	Max Score	Mean		SD		SMD	
				HS	AS	HS	AS	H1H2	HSAS
3	114749	1,363	1	0.45	0.44	0.50	0.50	-0.03	0.01
3	83640	1,008	2	0.83	0.88	0.75	0.76	0.00	-0.06
4	114768	1,401	1	0.36	0.35	0.48	0.48	-0.02	0.02

Grade	Item ID	N HSAS	Max Score	Mean		SD		SMD	
				HS	AS	HS	AS	H1H2	HSAS
4	91650	1,024	2	0.98	1.03	0.78	0.78	-0.00	-0.07
5	114786	1,280	1	0.55	0.56	0.50	0.50	-0.01	-0.03
5	84308	800	2	0.64	0.64	0.77	0.75	-0.01	-0.00
6	114807	1,774	1	0.64	0.64	0.48	0.48	0.01	0.00
6	2224	1,439	2	1.22	1.27	0.80	0.81	-0.00	-0.06
7	114822	1,782	1	0.40	0.40	0.49	0.49	0.00	-0.00
7	90459	1,541	2	1.15	1.17	0.79	0.79	0.00	-0.03
8	114840	1,955	1	0.67	0.67	0.47	0.47	0.00	-0.00
8	89173	1,933	2	1.17	1.18	0.76	0.76	0.00	-0.01
9	113231	3,462	1	0.45	0.45	0.50	0.50	-0.00	0.01
9	90632	2,161	2	1.40	1.41	0.72	0.71	-0.00	-0.01
10	113258	2,260	1	0.62	0.60	0.49	0.49	0.00	0.03
10	89405	1,919	2	1.26	1.30	0.74	0.75	0.00	-0.05
			Avg.					-0.00	-0.02

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses.

Table 13 presents ASE performance of ECR items with respect to exact agreement and QWK for each dimension. The performance of ASE for each dimension is above the target QWK threshold of .70.

Table 13. Performance of ASE with respect to Exact Agreement and QWK on ECR items in the held-out validation sample

Grade	Item ID	N HSAS	Dim.	Agreement			QWK
				HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS
3	12624	847	Conv.	61.6%	30.9%	7.4%	0.83
			Ideas	57.4%	35.5%	7.1%	0.85
4	12628	1,207	Conv.	54.4%	37.6%	8.0%	0.83
			Ideas	48.0%	41.4%	10.6%	0.88
5	12647	1,186	Conv.	62.1%	29.6%	8.3%	0.85
			Ideas	56.1%	32.9%	11.0%	0.88
6	12674	834	Conv.	58.6%	34.2%	7.2%	0.85
			Ideas	54.1%	36.6%	9.4%	0.91
7	61507	785	Conv.	57.6%	33.5%	8.9%	0.84
			Ideas	55.7%	37.3%	7.0%	0.91
8	73974	1,056	Conv.	59.5%	32.7%	7.9%	0.86
			Ideas	50.7%	42.5%	6.8%	0.91
9	68219	1,194	Conv.	62.8%	30.8%	6.4%	0.88
			Ideas	55.9%	37.9%	6.2%	0.93

Grade	Item ID	N HSAS	Dim.	Agreement			QWK
				HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS
10	69030	376	Conv.	57.7%	34.3%	8.0%	0.86
			Ideas	48.9%	41.0%	10.1%	0.89

Note: For ECR items, target performance for QWK is a value greater than 0.70. There are no target performance metrics for exact agreement. N HSAS is the number of human-scored responses.

Table 14 presents the HS and AS means and standard deviations, as well as the SMD values for ECR items. For all items, the SMD values were within performance thresholds.

Table 14. Performance of ASE with respect to SMD on the ECR items in the held-out validation sample

Grade	Item ID	N HSAS	Dim.	Max Score	Mean		SD		SMD	
					HS	AS	HS	AS	H1H2	HSAS
3	12624	847	Conv.	4	1.36	1.27	1.42	1.31	-0.01	0.06
			Ideas	6	1.86	1.85	1.57	1.47	-0.03	0.00
4	12628	1,207	Conv.	4	1.83	1.80	1.50	1.49	-0.01	0.02
			Ideas	6	2.56	2.60	1.95	1.86	-0.00	-0.02
5	12647	1,186	Conv.	4	1.31	1.30	1.53	1.47	0.01	0.01
			Ideas	6	1.83	1.67	1.97	1.83	0.01	0.08
6	12674	834	Conv.	4	1.56	1.44	1.53	1.52	0.02	0.08
			Ideas	6	2.40	2.28	2.15	2.10	-0.03	0.06
7	61507	785	Conv.	4	1.72	1.80	1.53	1.54	0.01	-0.06
			Ideas	6	2.32	2.33	1.95	1.99	0.01	-0.01
8	73974	1,056	Conv.	4	2.07	1.99	1.55	1.54	0.01	0.05
			Ideas	6	2.62	2.63	1.99	1.95	-0.01	-0.01
9	68219	1,194	Conv.	4	1.98	1.92	1.61	1.58	-0.01	0.03
			Ideas	6	2.68	2.73	2.17	2.09	0.01	-0.02
10	69030	376	Conv.	4	2.30	2.41	1.60	1.63	0.01	-0.07
			Ideas	6	3.04	3.08	2.04	2.08	0.00	-0.02

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses.

Table 15 and Table 16 present dimension-level agreement statistics for the models compared to the human raters on the rubric scale. Recall that the rubric-based scores are not reported, but rather are used to compute the summed dimension score. As such, statistics presented in these tables report do not reflect students' actual test scores or the overall performance of the automated scoring engine on reported scores. Still, the rubric scores do contribute to the summed score performance and it is important to evaluate human and ASE performance at this level.

Table 15 presents the Exact Agreement and Quadratic Weighted Kappa (QWK) of human-human agreement (H1H2), human-machine agreement (H1M1, H2M2), and the difference between the two, for each essay item. All models met performance criteria except for two exact agreement violations in Model 1 (the classical model).

Table 15. Performance of ASE compared to human-human agreement, with respect to Exact Agreement and QWK, for ECR items on the rubric dimensions in the held-out validation sample

Grade	Item ID	N	Dim.	H1H2	H1M1		H2M2		H1H2	H1M1		H2M2	
				EA	EA	diff	EA	diff	QWK	QWK	diff	QWK	diff
3	12624	847	Conv.	70.7%	68.7%	-2.0%	73.1%	2.4%	0.75	0.68	-0.07	0.75	0.00
3	12624	847	Ideas	70.1%	70.5%	0.4%	71.0%	0.8%	0.78	0.75	-0.04	0.78	-0.01
4	12628	1,207	Conv.	68.8%	63.1%	-5.6%	70.2%	1.4%	0.76	0.67	-0.09	0.75	-0.00
4	12628	1,207	Ideas	61.4%	61.1%	-0.3%	62.3%	0.9%	0.82	0.79	-0.03	0.80	-0.01
5	12647	1,186	Conv.	73.8%	71.1%	-2.7%	75.8%	2.0%	0.79	0.71	-0.09	0.79	0.00
5	12647	1,186	Ideas	67.3%	67.1%	-0.2%	70.6%	3.3%	0.84	0.81	-0.03	0.82	-0.02
6	12674	834	Conv.	70.6%	70.5%	-0.1%	73.1%	2.5%	0.78	0.76	-0.02	0.79	0.01
6	12674	834	Ideas	65.8%	65.5%	-0.4%	71.1%	5.3%	0.86	0.83	-0.03	0.87	0.01
7	61507	785	Conv.	68.5%	68.4%	-0.1%	69.3%	0.8%	0.76	0.72	-0.04	0.76	-0.00
7	61507	785	Ideas	65.0%	66.2%	1.3%	72.2%	7.3%	0.83	0.82	-0.01	0.86	0.02
8	73974	1,056	Conv.	69.8%	72.1%	2.3%	71.7%	1.9%	0.78	0.78	0.00	0.77	-0.00
8	73974	1,056	Ideas	63.9%	65.0%	1.0%	69.2%	5.3%	0.83	0.82	-0.01	0.85	0.02
9	68219	1,194	Conv.	75.9%	72.1%	-3.8%	76.2%	0.3%	0.83	0.79	-0.04	0.83	-0.00
9	68219	1,194	Ideas	69.6%	70.7%	1.1%	73.2%	3.6%	0.88	0.88	-0.00	0.88	0.00
10	69030	376	Conv.	69.9%	64.6%	-5.3%	73.4%	3.5%	0.79	0.74	-0.06	0.79	0.00
10	69030	376	Ideas	63.0%	59.8%	-3.2%	64.6%	1.6%	0.83	0.78	-0.05	0.84	0.01

Note: Target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10. H1M1 reflects the model 1 performance relative to rater 1. H2M2 refers to model 2 performance relative to rater 2.

Model 1 also shows one SMD violation and model 2 shows two SMD violations, all in the Conventions dimension (Table 16).

Table 16. Performance of ASE compared to human-human agreement, with respect to SMD, for ECR items on the rubric dimensions in the held-out validation sample

				Mean				SD				SMD		
Grade	Item ID	N	Dim.	H1	H2	M1	M2	H1	H2	M1	M2	H1H2	H1M1	H2M2
3	12624	847	Conv.	0.68	0.68	0.56	0.70	0.76	0.76	0.68	0.76	-0.01	0.16	-0.03
3	12624	847	Ideas	0.92	0.94	0.84	1.01	0.83	0.84	0.75	0.82	-0.03	0.10	-0.09
4	12628	1,207	Conv.	0.91	0.92	0.88	0.96	0.79	0.81	0.74	0.86	-0.01	0.04	-0.05
4	12628	1,207	Ideas	1.28	1.28	1.25	1.37	1.02	1.02	0.99	1.01	-0.00	0.02	-0.09
5	12647	1,186	Conv.	0.66	0.65	0.56	0.71	0.81	0.81	0.78	0.84	0.01	0.13	-0.07
5	12647	1,186	Ideas	0.92	0.91	0.79	0.83	1.02	1.03	1.00	0.96	0.01	0.13	0.08
6	12674	834	Conv.	0.79	0.77	0.70	0.78	0.82	0.81	0.79	0.89	0.02	0.11	-0.00

				Mean				SD				SMD		
Grade	Item ID	N	Dim.	H1	H2	M1	M2	H1	H2	M1	M2	H1H2	H1M1	H2M2
6	12674	834	Ideas	1.19	1.22	1.08	1.18	1.10	1.12	1.10	1.09	-0.03	0.09	0.03
7	61507	785	Conv.	0.86	0.85	0.77	1.03	0.81	0.82	0.81	0.88	0.01	0.11	-0.21
7	61507	785	Ideas	1.17	1.15	1.08	1.25	1.03	1.01	1.00	1.08	0.01	0.09	-0.09
8	73974	1,056	Conv.	1.04	1.03	1.01	0.99	0.83	0.82	0.82	0.83	0.01	0.03	0.05
8	73974	1,056	Ideas	1.30	1.31	1.28	1.34	1.03	1.05	1.02	1.03	-0.01	0.03	-0.03
9	68219	1,194	Conv.	0.98	0.99	0.97	0.99	0.84	0.85	0.84	0.86	-0.01	0.02	0.00
9	68219	1,194	Ideas	1.35	1.34	1.31	1.41	1.12	1.12	1.12	1.06	0.01	0.04	-0.07
10	69030	376	Conv.	1.15	1.14	1.16	1.28	0.85	0.85	0.86	0.81	0.01	-0.01	-0.17
10	69030	376	Ideas	1.52	1.52	1.47	1.63	1.07	1.06	1.05	1.15	0.00	0.05	-0.10

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses. H1 refers to human rater 1. H2 refers to human rater 2. M1 refers to model 1. M2 refers to model 2.

Random Percent Routing

ASE performance on the Random Percent sample included aggregate performance for each item, as well as within-group performance by gender (male, female), race/ethnicity (Black, Latino, White), economically disadvantaged status (Eco-disc), and emergent bilingual (EB) students. For score point distributions for human and ASE models, please see Appendix B. For item-level information on student group performance, please see Appendix C.

ASE Performance in the Aggregate

Table 17 presents the H1H2 and HSAS exact agreements and QWK values for each SCR item. Across all items and measures, values were similar between H1H2 and HSAS. All items met the evaluation criteria.

Table 17. Performance of ASE with respect to Exact Agreement and QWK on SCR items in the random percent sample

Grade	Item ID	N HSAS	N H1H2	Max Score	Exact Agreement			QWK		
					H1H2	HSAS	diff	H1H2	HSAS	diff
3	114749	33,514	6,138	1	96.3%	97.2%	0.9%	0.92	0.94	0.02
3	83640	34,273	7,056	2	77.9%	79.6%	1.7%	0.79	0.81	0.02
4	114768	35,110	6,435	1	96.1%	97.3%	1.2%	0.91	0.94	0.03
4	91650	36,482	10,315	2	67.9%	72.4%	4.5%	0.68	0.74	0.06
5	114786	36,724	6,334	1	91.2%	93.4%	2.3%	0.82	0.87	0.04
5	84308	37,194	10,608	2	71.5%	75.0%	3.5%	0.73	0.77	0.05
6	114807	38,393	7,438	1	91.6%	94.5%	2.9%	0.82	0.88	0.06
6	2224	38,821	10,091	2	75.6%	81.4%	5.8%	0.78	0.84	0.05
7	114822	38,537	6,868	1	96.8%	97.9%	1.1%	0.93	0.96	0.02
7	90459	39,190	9,954	2	73.8%	77.6%	3.8%	0.75	0.80	0.04
8	114840	39,303	6,971	1	89.7%	93.5%	3.8%	0.78	0.86	0.08

Grade	Item ID	N HSAS	N H1H2	Max Score	Exact Agreement			QWK		
					H1H2	HSAS	diff	H1H2	HSAS	diff
8	89173	39,478	10,856	2	76.0%	80.7%	4.7%	0.77	0.82	0.05
9	113231	45,861	8,992	1	97.2%	98.2%	1.0%	0.94	0.96	0.02
9	90632	47,355	10,334	2	79.3%	82.0%	2.7%	0.80	0.82	0.02
10	113258	44,213	8,918	1	92.0%	93.6%	1.6%	0.84	0.87	0.03
10	89405	45,037	11,878	2	84.1%	79.5%	-4.6%	0.85	0.81	-0.04
				Avg.	84.8%	87.1%	2.3%	0.82	0.86	0.04

Note: For SCR items, target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than .10. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

Table 18 presents the HS and AS means and standard deviations, as well as the SMD values for SCR items. For all items, the SMD values were within performance thresholds. Grade 3 item 83640 was close to the threshold, with an SMD magnitude of .13.

Table 18. Performance of ASE with respect to SMD on SCR items in the random percent sample

Grade	Item ID	N HSAS	Max Score	Mean		SD		SMD	
				HS	AS	HS	AS	H1H2	HSAS
3	114749	33,514	1	0.45	0.45	0.50	0.50	-0.01	-0.01
3	83640	34,273	2	0.81	0.91	0.75	0.76	0.01	-0.13
4	114768	35,110	1	0.35	0.35	0.48	0.48	0.00	0.01
4	91650	36,482	2	1.02	1.03	0.75	0.76	0.00	-0.02
5	114786	36,724	1	0.57	0.58	0.49	0.49	-0.00	-0.01
5	84308	37,194	2	0.74	0.75	0.78	0.78	0.00	-0.01
6	114807	38,393	1	0.65	0.65	0.48	0.48	-0.00	0.00
6	2224	38,821	2	1.22	1.27	0.79	0.79	0.02	-0.06
7	114822	38,537	1	0.41	0.41	0.49	0.49	-0.00	-0.01
7	90459	39,190	2	1.22	1.22	0.77	0.77	0.01	-0.00
8	114840	39,303	1	0.65	0.67	0.48	0.47	-0.01	-0.04
8	89173	39,478	2	1.18	1.19	0.75	0.74	0.01	-0.02
9	113231	45,861	1	0.45	0.45	0.50	0.50	-0.00	0.01
9	90632	47,355	2	1.36	1.39	0.73	0.70	0.00	-0.04
10	113258	44,213	1	0.58	0.59	0.49	0.49	0.03	-0.03
10	89405	45,037	2	1.30	1.27	0.74	0.73	-0.01	0.04
			Avg.					0.00	-0.02

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

Table 19 presents ASE performance of ECR items with respect to exact agreement and QWK. The performance of ASE for each dimension is above the target QWK threshold of .70. Performance is uniformly good across all items and dimensions relative to the QWK metric.

Table 19. Performance of ASE with respect to Exact Agreement and QWK on ECR items in the random percent sample

Grade	Item ID	N HSAS	Dim.	Agreement			QWK
				HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS
3	12624	29,625	Conv.	57.3%	35.2%	7.5%	0.83
			Ideas	58.9%	33.9%	7.2%	0.87
4	12628	31,955	Conv.	54.4%	36.9%	8.7%	0.82
			Ideas	49.3%	40.1%	10.6%	0.87
5	12647	34,128	Conv.	60.4%	30.0%	9.6%	0.83
			Ideas	53.7%	34.7%	11.6%	0.86
6	12674	35,671	Conv.	56.1%	32.4%	11.5%	0.81
			Ideas	52.4%	34.7%	12.9%	0.88
7	61507	36,722	Conv.	55.7%	35.5%	8.8%	0.83
			Ideas	53.8%	37.9%	8.2%	0.90
8	73974	36,407	Conv.	59.1%	33.9%	7.0%	0.86
			Ideas	53.7%	38.4%	7.9%	0.90
9	68219	42,086	Conv.	61.1%	31.9%	7.0%	0.88
			Ideas	59.0%	34.0%	7.0%	0.92
10	69030	41,233	Conv.	58.8%	29.7%	11.5%	0.82
			Ideas	49.8%	38.0%	12.2%	0.87

Note: For ECR items, target performance for QWK is a value greater than 0.70. There are no target performance metrics for exact agreement. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

Table 20 presents the HS and AS means and standard deviations, as well as the SMD values for ECR items. For all items, the SMD values were within performance thresholds.

Table 20. Performance of ASE with respect to SMD on the ECR items in the random percent sample

Grade	Item ID	N HSAS	Dim.	Max Score	Mean		SD		SMD	
					HS	AS	HS	AS	H1H2	HSAS
3	12624	29,625	Conv.	4	1.54	1.39	1.49	1.37	-0.01	0.10
			Ideas	6	1.93	1.99	1.59	1.59	-0.00	-0.03
4	12628	31,955	Conv.	4	1.79	1.88	1.49	1.49	-0.00	-0.06
			Ideas	6	2.51	2.63	1.92	1.86	0.00	-0.07
5	12647	34,128	Conv.	4	1.27	1.33	1.54	1.47	0.01	-0.04

					Mean		SD		SMD	
Grade	Item ID	N HSAS	Dim.	Max Score	HS	AS	HS	AS	H1H2	HSAS
			Ideas	6	1.73	1.75	1.99	1.85	0.01	-0.01
6	12674	35,671	Conv.	4	1.64	1.59	1.59	1.56	-0.01	0.04
			Ideas	6	2.57	2.52	2.26	2.12	0.00	0.02
7	61507	36,722	Conv.	4	1.93	2.03	1.51	1.51	0.01	-0.07
			Ideas	6	2.57	2.66	1.90	1.96	0.01	-0.05
8	73974	36,407	Conv.	4	2.16	2.06	1.57	1.51	-0.00	0.06
			Ideas	6	2.68	2.70	1.99	1.91	-0.00	-0.01
9	68219	42,086	Conv.	4	1.92	1.80	1.62	1.56	0.00	0.08
			Ideas	6	2.47	2.56	2.12	2.09	0.00	-0.04
10	69030	41,233	Conv.	4	2.22	2.31	1.63	1.62	-0.01	-0.06
			Ideas	6	2.80	2.91	2.04	2.08	0.00	-0.05

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

As was done for the held-out validation sample, the performance can be evaluated at the rubric level for each dimension. Again, these results do not reflect scores as reported to students but do provide insight into the rubric-level scoring.

Table 21 presents the Exact Agreement and Quadratic Weighted Kappa (QWK) of human-human agreement (H1H2), human-machine agreement (H1M1 and H2M2), and the difference between the two, for each ECR item. All models met the performance criteria.

Table 21. Performance of ASE compared to human-human agreement on the ECR dimension rubric scores, with respect to Exact Agreement and QWK in the random percent sample

				H1H2	H1M1		H2M2		H1H2	H1M1		H2M2	
Grade	Item ID	N	Dim.	EA	EA	diff	EA	diff	QWK	QWK	diff	QWK	diff
3	12624	29,625	Conv.	69.2%	66.9%	-2.3%	73.3%	4.1%	0.71	0.68	-0.03	0.77	0.06
3	12624	29,625	Ideas	67.8%	69.9%	2.1%	73.6%	5.7%	0.74	0.76	0.02	0.80	0.06
4	12628	31,955	Conv.	66.1%	63.3%	-2.9%	69.8%	3.7%	0.67	0.67	-0.01	0.73	0.06
4	12628	31,955	Ideas	59.6%	60.5%	0.8%	66.3%	6.6%	0.73	0.76	0.03	0.79	0.06
5	12647	34,128	Conv.	73.5%	70.5%	-3.1%	74.1%	0.6%	0.71	0.70	-0.01	0.76	0.05
5	12647	34,128	Ideas	68.7%	65.6%	-3.1%	71.4%	2.6%	0.77	0.78	0.01	0.80	0.03
6	12674	35,671	Conv.	66.9%	66.7%	-0.2%	70.5%	3.6%	0.67	0.69	0.02	0.74	0.07
6	12674	35,671	Ideas	63.4%	64.0%	0.5%	68.2%	4.7%	0.78	0.82	0.03	0.84	0.05
7	61507	36,722	Conv.	65.8%	66.3%	0.5%	68.2%	2.5%	0.69	0.71	0.02	0.74	0.04
7	61507	36,722	Ideas	64.0%	67.4%	3.3%	67.7%	3.7%	0.79	0.82	0.03	0.83	0.04
8	73974	36,407	Conv.	68.7%	68.7%	0.0%	74.0%	5.2%	0.73	0.75	0.02	0.79	0.06

				H1H2	H1M1		H2M2		H1H2	H1M1		H2M2	
Grade	Item ID	N	Dim.	EA	EA	diff	EA	diff	QWK	QWK	diff	QWK	diff
8	73974	36,407	Ideas	61.5%	65.8%	4.4%	69.0%	7.5%	0.78	0.82	0.04	0.84	0.06
9	68219	42,086	Conv.	73.5%	69.5%	-4.0%	75.9%	2.4%	0.78	0.76	-0.01	0.81	0.04
9	68219	42,086	Ideas	68.1%	69.8%	1.7%	72.4%	4.3%	0.84	0.86	0.02	0.87	0.03
10	69030	41,233	Conv.	67.8%	66.1%	-1.7%	70.9%	3.1%	0.69	0.71	0.02	0.74	0.05
10	69030	41,233	Ideas	59.6%	62.3%	2.6%	63.9%	4.2%	0.75	0.79	0.04	0.81	0.05

Note: Target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than .10. H1M1 reflects the model 1 performance relative to rater 1. H2M2 refers to model 2 performance relative to rater 2.

Table 22 presents the means, standard deviations, and Standardized Mean Difference (SMD) of human-human agreement (H1H2) and human-machine agreement (H1M1 and H2M2), for each item and dimension. There are two SMD violations at the dimension level, both in Conventions and one for each model

Table 22. Performance of ASE compared to human-human agreement, with respect to SMD, on the ECR dimension rubric level scores

				Mean				SD				SMD		
Grade	Item ID	N	Dim.	H1	H2	M1	M2	H1	H2	M1	M2	H1H2	H1M1	H2M2
3	12624	29,625	Conv.	0.78	0.79	0.63	0.76	0.80	0.80	0.73	0.78	-0.01	0.20	0.03
3	12624	29,625	Ideas	0.98	0.98	0.92	1.07	0.85	0.85	0.81	0.86	-0.00	0.07	-0.11
4	12628	31,955	Conv.	0.89	0.89	0.92	1.00	0.80	0.80	0.76	0.86	-0.00	-0.04	-0.13
4	12628	31,955	Ideas	1.26	1.25	1.27	1.37	1.01	1.02	0.99	1.00	0.00	-0.01	-0.11
5	12647	34,128	Conv.	0.65	0.65	0.59	0.72	0.82	0.81	0.80	0.83	0.01	0.07	-0.08
5	12647	34,128	Ideas	0.88	0.87	0.84	0.87	1.03	1.03	1.00	0.98	0.01	0.04	0.00
6	12674	35,671	Conv.	0.84	0.84	0.77	0.84	0.84	0.84	0.82	0.89	-0.01	0.09	0.01
6	12674	35,671	Ideas	1.31	1.31	1.22	1.30	1.16	1.16	1.11	1.10	0.00	0.08	0.01
7	61507	36,722	Conv.	0.97	0.96	0.90	1.15	0.81	0.81	0.82	0.83	0.01	0.09	-0.23
7	61507	36,722	Ideas	1.28	1.28	1.25	1.41	0.99	1.00	1.00	1.06	0.01	0.04	-0.13
8	73974	36,407	Conv.	1.08	1.09	1.06	1.02	0.83	0.83	0.81	0.82	-0.00	0.04	0.08
8	73974	36,407	Ideas	1.35	1.35	1.31	1.37	1.05	1.05	1.00	1.01	-0.00	0.03	-0.03
9	68219	42,086	Conv.	0.96	0.96	0.93	0.91	0.85	0.85	0.83	0.86	0.00	0.04	0.06
9	68219	42,086	Ideas	1.24	1.24	1.22	1.33	1.10	1.10	1.11	1.07	0.00	0.02	-0.09
10	69030	41,233	Conv.	1.11	1.11	1.11	1.24	0.86	0.87	0.84	0.82	-0.01	-0.01	-0.15
10	69030	41,233	Ideas	1.40	1.40	1.40	1.52	1.07	1.07	1.04	1.13	0.00	-0.00	-0.11

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

ASE Performance by Student Group

It is important to ensure that ASE is performing well, not just overall, but for student groups. In this section, we analyze ASE performance, disaggregated by student group. Specifically, we examine performance across female and male students, Black, Latino, and White students, students with economically disadvantaged status (Eco-disc) and emergent bilingual (EB) students. We begin these analyses by presenting numbers and percentages of each student group for each item (Table 23).

Table 23. Distribution of responses in the random percent sample, by student group

Grade	Item ID	Item Type	Max Score	All	Female		Male		Black		Latino		White		Eco-disc		EB	
				N	N	%	N	%	N	%	N	%	N	%	N	%	N	%
3	114749	SCR	1	33,514	16,798	50.1%	16,705	49.8%	4,402	13.1%	16,085	48.0%	9,281	27.7%	19,796	59.1%	6,540	19.5%
3	83640	SCR	2	34,273	17,094	49.9%	17,172	50.1%	4,536	13.2%	16,472	48.1%	9,402	27.4%	20,333	59.3%	6,855	20.0%
3	12624	ECR	4,6	29,625	15,188	51.3%	14,426	48.7%	3,682	12.4%	14,159	47.8%	8,338	28.1%	16,926	57.1%	5,949	20.1%
4	114768	SCR	1	35,110	17,522	49.9%	17,580	50.1%	4,462	12.7%	17,304	49.3%	9,382	26.7%	20,756	59.1%	7,569	21.6%
4	91650	SCR	2	36,482	17,866	49.0%	18,609	51.0%	4,776	13.1%	18,063	49.5%	9,760	26.8%	21,912	60.1%	8,179	22.4%
4	12628	ECR	4,6	31,955	16,124	50.5%	15,825	49.5%	3,865	12.1%	15,554	48.7%	8,875	27.8%	18,458	57.8%	6,917	21.6%
5	114786	SCR	1	36,724	18,163	49.5%	18,550	50.5%	4,686	12.8%	18,325	49.9%	9,786	26.6%	21,974	59.8%	8,620	23.5%
5	84308	SCR	2	37,194	18,451	49.6%	18,737	50.4%	4,828	13.0%	18,621	50.1%	9,843	26.5%	22,434	60.3%	8,667	23.3%
5	12647	ECR	4,6	34,128	17,138	50.2%	16,980	49.8%	4,233	12.4%	16,989	49.8%	9,250	27.1%	20,181	59.1%	7,931	23.2%
6	114807	SCR	1	38,393	19,099	49.7%	19,280	50.2%	4,752	12.4%	19,910	51.9%	9,761	25.4%	23,201	60.4%	10,036	26.1%
6	2224	SCR	2	38,821	19,072	49.1%	19,742	50.9%	4,827	12.4%	20,240	52.1%	9,821	25.3%	23,793	61.3%	10,163	26.2%
6	12674	ECR	4,6	35,671	17,882	50.1%	17,776	49.8%	4,278	12.0%	18,322	51.4%	9,403	26.4%	21,154	59.3%	8,824	24.7%
7	114822	SCR	1	38,537	19,117	49.6%	19,413	50.4%	4,793	12.4%	20,153	52.3%	9,707	25.2%	23,022	59.7%	9,834	25.5%
7	90459	SCR	2	39,190	19,132	48.8%	20,041	51.1%	4,933	12.6%	20,687	52.8%	9,619	24.5%	23,623	60.3%	10,177	26.0%
7	61507	ECR	4,6	36,722	18,358	50.0%	18,352	50.0%	4,465	12.2%	18,963	51.6%	9,553	26.0%	21,653	59.0%	8,936	24.3%
8	114840	SCR	1	39,303	19,236	48.9%	20,051	51.0%	4,873	12.4%	20,547	52.3%	10,000	25.4%	23,443	59.6%	9,570	24.3%
8	89173	SCR	2	39,478	19,346	49.0%	20,119	51.0%	4,824	12.2%	20,719	52.5%	10,053	25.5%	23,587	59.7%	9,645	24.4%
8	73974	ECR	4,6	36,407	18,412	50.6%	17,990	49.4%	4,436	12.2%	18,782	51.6%	9,564	26.3%	21,230	58.3%	8,346	22.9%
9	113231	SCR	1	45,861	22,387	48.8%	23,464	51.2%	6,093	13.3%	24,975	54.5%	10,739	23.4%	28,270	61.6%	11,900	25.9%
9	90632	SCR	2	47,355	22,488	47.5%	24,858	52.5%	6,236	13.2%	26,067	55.0%	10,794	22.8%	29,524	62.3%	12,704	26.8%
9	68219	ECR	4,6	42,086	20,846	49.5%	21,234	50.5%	5,366	12.8%	22,707	54.0%	10,061	23.9%	25,271	60.0%	10,117	24.0%
10	113258	SCR	1	44,213	21,677	49.0%	22,528	51.0%	5,680	12.8%	24,147	54.6%	10,417	23.6%	26,414	59.7%	10,228	23.1%
10	89405	SCR	2	45,037	21,782	48.4%	23,243	51.6%	5,674	12.6%	24,403	54.2%	10,847	24.1%	26,818	59.5%	10,508	23.3%
10	69030	ECR	4,6	41,233	20,464	49.6%	20,763	50.4%	5,073	12.3%	22,201	53.8%	10,172	24.7%	24,075	58.4%	8,811	21.4%

Grade	Item ID	Item Type	Max Score	All	Female		Male		Black		Latino		White		Eco-disc		EB	
				N	N	%	N	%	N	%	N	%	N	%	N	%	N	%
			Total	917,312	453,642	49.0%	463,438	50.0%	115,773	12.5%	474,395	51.2%	234,428	25.3%	547,848	59.1%	217,026	23.4%

In Table 24, we present average performance within item types, for each student group. Using the same performance thresholds used on the aggregate data, we examine whether items meet the criteria within student group. For 1-point SCRs and ECRs, all items for all student groups meet all three criteria. For 2-point SCRs, all items met the criteria for female students, Black, Latino, and White students, students with Eco-disc backgrounds, and EB students. One item did not meet the criteria for males and for EB students. Appendix C presents the individual item results for each group.

Table 24. Overall performance of ASE with respect to Exact Agreement, QWK, and SMD in the random percent sample, disaggregated by student group

Item Type	Dim.	Max Score	Std. Group	Exact Agreement			QWK			SMD			Comb.
				H1H2	HSAS	Meets	H1H2	HSAS	Meets	H1H2	HSAS	Meets	Meets
SCR	Overall	1	Female	94.1%	95.8%	100%	0.87	0.91	100%	-0.00	-0.01	100%	100%
			Male	93.6%	95.6%	100%	0.87	0.91	100%	0.00	-0.01	100%	100%
			Black	94.1%	95.4%	100%	0.87	0.90	100%	-0.01	-0.00	100%	100%
			Latino	93.7%	95.7%	100%	0.87	0.91	100%	0.00	-0.01	100%	100%
			White	94.0%	95.7%	100%	0.86	0.90	100%	0.00	-0.01	100%	100%
			Eco-disc	93.7%	95.6%	100%	0.87	0.91	100%	0.00	-0.01	100%	100%
			EB	93.6%	95.6%	100%	0.86	0.91	100%	0.00	-0.01	100%	100%
SCR	Overall	2	Female	75.3%	78.5%	100%	0.75	0.79	100%	0.01	-0.05	100%	100%
			Male	76.2%	78.6%	88%	0.78	0.81	100%	0.00	-0.02	100%	88%
			Black	75.5%	78.8%	100%	0.76	0.80	100%	-0.00	-0.02	100%	100%
			Latino	75.6%	78.4%	100%	0.77	0.80	100%	0.00	-0.02	100%	100%
			White	75.7%	78.1%	100%	0.76	0.78	100%	0.01	-0.04	100%	100%
			Eco-disc	75.6%	78.5%	100%	0.76	0.80	100%	0.01	-0.02	100%	100%
			EB	75.6%	78.4%	88%	0.77	0.80	100%	0.01	-0.02	100%	88%
ECR	Conv.	4	Female		56.3%			0.83	100%	-0.01	0.00	100%	100%
			Male		59.5%			0.84	100%	0.00	0.01	100%	100%
			Black		59.9%			0.83	100%	0.00	0.01	100%	100%
			Latino		58.5%			0.83	100%	-0.00	-0.00	100%	100%
			White		55.4%			0.82	100%	-0.00	0.03	100%	100%
			Eco-disc		59.5%			0.83	100%	-0.00	0.00	100%	100%
			EB		60.0%			0.82	100%	0.01	-0.01	100%	100%
ECR	Ideas	6	Female		52.0%			0.88	100%	-0.00	-0.04	100%	100%
			Male		55.7%			0.89	100%	0.01	-0.03	100%	100%
			Black		56.8%			0.88	100%	0.00	-0.03	100%	100%
			Latino		55.0%			0.88	100%	0.00	-0.03	100%	100%
			White		50.8%			0.87	100%	0.00	-0.02	100%	100%

Item Type	Dim.	Max Score	Std. Group	Exact Agreement			QWK			SMD			Comb.
				H1H2	HSAS	Meets	H1H2	HSAS	Meets	H1H2	HSAS	Meets	Meets
			Eco-disc		56.2%			0.88	100%	0.00	-0.03	100%	100%
			EB		57.3%			0.87	100%	0.01	-0.04	100%	100%

Note: Meets indicates the percentage of items that reached target performance. Combined (Comb.) Meets indicates percentage of items that reached target performance on all three metrics.

Condition Code Routing

Here we present results on responses routed due to receiving an ASE condition code that is flagged for routing to human raters. Recall that both models can influence the routing process for condition codes. In this section, we present only condition codes routed by the reprogrammed model. Note that if responses were routed under the original model due to a condition code, the human score serves as the final score of record. Additionally, because most condition codes are identified using the algorithmic condition codes not routed for human scoring, the NONSPECIFIC model was often not programmed for items; this was because this model is programmed on human-assigned condition codes appearing in the data. It turned out that there were very few human-assigned condition codes in the reprogramming sample, and too few to program on this model.

All routed condition codes were scored by expert human raters. Table 26 presents the distribution of scores for each item, disaggregated by condition codes by the final model. Out of Vocabulary CCs were generally scored as 0s by human raters, at a rate of 97-100%. All responses identified as Nonspecific were scored as 0s for grade 3 but showed a range of higher scores for grade 4. Unusual scores, as expected, received a range of rubric scores by the expert reads; this code routes responses for which ASE gave a rubric score that was either higher than expected or for which the two models provided non-adjacent scores.

Table 25. Score point distributions of human rater scores, by routed condition code

Condition Code	Grade	Item ID	Item Type	N	Dim.	Max Score	Score Point Distribution						
							0	1	2	3	4	5	6
Out of Vocab.	3	114749	SCR	6,838	Overall	1	100	0					
	3	83640	SCR	12,618	Overall	2	99	1	0				
	3	12624	ECR	1,868	Conv.	4	99	1	0	0	0		
	3	12624	ECR	1,868	Ideas	6	97	2	1	0	0	0	0
	4	114768	SCR	4,635	Overall	1	100	0					
	4	91650	SCR	4,670	Overall	2	99	1	0				
	4	12628	ECR	896	Conv.	4	100	0	0	0	0		
	4	12628	ECR	896	Ideas	6	100	0	0	0	0	0	0
	5	114786	SCR	2,663	Overall	1	99	1					
	5	84308	SCR	3,425	Overall	2	99	1	0				
	5	12647	ECR	629	Conv.	4	100	0	0	0	0		
	5	12647	ECR	629	Ideas	6	100	0	0	0	0	0	0

							Score Point Distribution						
Condition Code	Grade	Item ID	Item Type	N	Dim.	Max Score	0	1	2	3	4	5	6
	6	114807	SCR	2,340	Overall	1	100	0					
	6	2224	SCR	3,295	Overall	2	99	1	0				
	6	12674	ECR	851	Conv.	4	100	0	0	0	0		
	6	12674	ECR	851	Ideas	6	100	0	0	0	0	0	0
	7	114822	SCR	1,732	Overall	1	100	0					
	7	90459	SCR	2,293	Overall	2	99	1	0				
	7	61507	ECR	383	Conv.	4	99	0	1	0	0		
	7	61507	ECR	383	Ideas	6	98	0	1	0	0	0	0
	8	114840	SCR	1,230	Overall	1	100	0					
	8	89173	SCR	2,704	Overall	2	99	1	0				
	8	73974	ECR	314	Conv.	4	99	0	1	0	0		
	8	73974	ECR	314	Ideas	6	99	0	0	0	0	0	0
	9	113231	SCR	1,945	Overall	1	100	0					
	9	90632	SCR	3,930	Overall	2	100	0	0				
	9	68219	ECR	372	Conv.	4	100	0	0	0	0		
	9	68219	ECR	372	Ideas	6	100	0	0	0	0	0	0
	10	113258	SCR	1,484	Overall	1	100	0					
	10	89405	SCR	3,675	Overall	2	100	0	0				
	10	69030	ECR	279	Conv.	4	99	0	1	0	0		
	10	69030	ECR	279	Ideas	6	99	0	1	0	0	0	0
Unusual Scores	3	12624	ECR	5,025	Conv.	4	11	12	32	26	19		
	3	12624	ECR	5,025	Ideas	6	8	7	56	19	8	1	1
	4	12628	ECR	5,552	Conv.	4	33	19	30	12	7		
	4	12628	ECR	5,552	Ideas	6	27	14	27	17	11	3	1
	5	12647	ECR	7,496	Conv.	4	36	13	19	17	15		
	5	12647	ECR	7,496	Ideas	6	33	10	23	16	13	3	2
	6	12674	ECR	2,075	Conv.	4	54	12	13	11	10		
	6	12674	ECR	2,075	Ideas	6	48	7	8	9	14	7	7
	7	61507	ECR	1,367	Conv.	4	16	19	28	23	14		
	7	61507	ECR	1,367	Ideas	6	12	14	49	13	11	2	0
	8	73974	ECR	245	Conv.	4	28	24	26	13	9		
	8	73974	ECR	245	Ideas	6	22	18	31	15	9	2	2
	9	68219	ECR	2,308	Conv.	4	37	13	27	11	12		
	9	68219	ECR	2,308	Ideas	6	34	12	39	8	6	1	1
	10	69030	ECR	5,621	Conv.	4	38	8	23	10	22		
	10	69030	ECR	5,621	Ideas	6	34	7	35	7	14	2	1
Nonspecific	3	83640	SCR	5	Overall	2	100	0	0				

							Score Point Distribution						
Condition Code	Grade	Item ID	Item Type	N	Dim.	Max Score	0	1	2	3	4	5	6
	3	12624	ECR	1	Conv.	4	100	0	0	0	0		
	3	12624	ECR	1	Ideas	6	100	0	0	0	0	0	0
	4	12628	ECR	532	Conv.	4	54	14	16	8	9		
	4	12628	ECR	532	Ideas	6	49	12	10	10	13	5	2

Low Confidence Routing

Responses with confidence percentile values lower than the 10th percentile from either the original or reprogrammed model were routed for human scoring. This low confidence sample includes responses that were routed from either the original or reprogrammed models; however, note that the automated scores presented below are from the reprogrammed models. We do find that responses that were low confidence on the original model tend to have lower confidence values in the reprogrammed model, even if those low confidence values are not below the 10th percentile threshold. Regardless, we expect the engine agreements to be lower than human agreements on this sample because this threshold was set to capture the responses with which the engine predicts scores that are more likely to differ from scores assigned by human raters. Finally, note that score point distributions of human rater and machine scores from the low confidence sample are presented in Appendix D.

Table 26 presents Exact Agreement and QWK of HSAS compared to H1H2 for SCR items. Agreements are lower both among human raters and ASE, as compared to the random percent sample. On average, the two human raters agreed at 6% lower exact agreement rates on the low confidence sample compared to the random percent sample; ASE agreed at 15% lower rates on this sample. With regard to QWK, the human agreements were .13 lower and the engine agreement rates were .25 lower. Most of the items in the low confidence sample do not meet target performance, particularly with respect to Exact Agreement.

Table 26. Performance of ASE with respect to Exact Agreement and QWK on SCR items in the low confidence sample

					Exact Agreement			QWK		
Grade	Item ID	N HSAS	N H1H2	Max Score	H1H2	HSAS	diff	H1H2	HSAS	diff
3	114749	43,402	10,016	1	92.2%	85.4%	-6.8%	0.84	0.71	-0.13
3	83640	63,833	13,792	2	70.7%	62.0%	-8.7%	0.58	0.52	-0.06
4	114768	54,393	13,780	1	94.8%	89.2%	-5.7%	0.84	0.76	-0.08
4	91650	63,919	14,318	2	61.7%	56.7%	-5.0%	0.54	0.51	-0.03
5	114786	68,051	18,530	1	85.3%	79.5%	-5.8%	0.70	0.58	-0.12
5	84308	71,441	15,984	2	61.0%	56.8%	-4.2%	0.48	0.46	-0.02
6	114807	63,103	16,380	1	83.1%	80.8%	-2.3%	0.60	0.55	-0.05
6	2224	83,019	19,731	2	72.5%	67.0%	-5.5%	0.70	0.65	-0.05

Grade	Item ID	N HSAS	N H1H2	Max Score	Exact Agreement			QWK		
					H1H2	HSAS	diff	H1H2	HSAS	diff
7	114822	59,890	16,583	1	90.2%	90.4%	0.2%	0.80	0.80	0.00
7	90459	69,562	16,783	2	66.4%	55.7%	-10.7%	0.59	0.44	-0.14
8	114840	63,580	17,887	1	84.2%	82.3%	-2.0%	0.68	0.64	-0.03
8	89173	62,126	14,107	2	68.2%	61.2%	-6.9%	0.62	0.55	-0.06
9	113231	75,321	20,187	1	94.6%	95.1%	0.4%	0.89	0.89	0.00
9	90632	80,524	21,185	2	77.8%	63.9%	-13.9%	0.75	0.59	-0.16
10	113258	62,885	17,138	1	83.5%	75.0%	-8.5%	0.66	0.49	-0.18
10	89405	71,885	17,168	2	76.6%	59.6%	-17.0%	0.73	0.56	-0.17
				Avg.	78.9%	72.5%	-6.4%	0.69	0.61	-0.08

Note: For SCR items, target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10. N HSAS is the number of human-scored responses, whereas N H1H2 is the number of double-scored responses.

Table 27 presents SMD of SCR items for the low confidence sample. Here, too, there are violations of SMD.

Table 27. Performance of ASE with respect to SMD on the SCR items in the low confidence sample

Grade	Item ID	N HSAS	Max Score	Mean		SD		SMD	
				HS	AS	HS	AS	H1H2	HSAS
3	114749	43,402	1	0.53	0.57	0.50	0.49	-0.00	-0.10
3	83640	63,833	2	1.15	1.34	0.64	0.63	-0.00	-0.30
4	114768	54,393	1	0.34	0.33	0.47	0.47	0.00	0.02
4	91650	63,919	2	1.07	1.08	0.67	0.70	0.00	-0.02
5	114786	68,051	1	0.46	0.42	0.50	0.49	0.00	0.09
5	84308	71,441	2	0.86	0.89	0.70	0.63	0.01	-0.04
6	114807	63,103	1	0.30	0.31	0.46	0.46	-0.00	-0.02
6	2224	83,019	2	1.22	1.32	0.75	0.71	0.00	-0.14
7	114822	59,890	1	0.40	0.41	0.49	0.49	-0.00	-0.03
7	90459	69,562	2	1.29	1.07	0.66	0.65	-0.00	0.33
8	114840	63,580	1	0.40	0.49	0.49	0.50	-0.00	-0.19
8	89173	62,126	2	1.05	1.03	0.68	0.66	0.00	0.03
9	113231	75,321	1	0.38	0.38	0.49	0.48	0.00	0.00
9	90632	80,524	2	1.14	1.26	0.68	0.66	0.01	-0.18
10	113258	62,885	1	0.36	0.45	0.48	0.50	0.01	-0.18
10	89405	71,885	2	1.41	1.17	0.67	0.65	-0.00	0.37
			Avg.					0.00	-0.02

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses.

Similar to SCR items, Table 28 shows that ECR items also display lower Exact Agreement and lower QWK, most of which fall below the .70 threshold.

Table 28. Performance of ASE with respect to Exact Agreement and QWK on ECR items in the low confidence sample

Grade	Item ID	N HSAS	Dim.	Agreement			QWK
				HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS
3	12624	53,564	Conv.	43.5%	47.4%	9.1%	0.67
			Ideas	49.9%	39.9%	10.3%	0.69
4	12628	45,565	Conv.	39.2%	45.9%	14.8%	0.45
			Ideas	34.9%	44.1%	21.1%	0.51
5	12647	66,749	Conv.	34.4%	45.7%	19.9%	0.54
			Ideas	33.1%	44.1%	22.8%	0.65
6	12674	69,199	Conv.	42.4%	38.8%	18.7%	0.49
			Ideas	33.5%	37.3%	29.3%	0.62
7	61507	96,136	Conv.	41.2%	46.7%	12.1%	0.61
			Ideas	40.6%	47.1%	12.3%	0.72
8	73974	75,873	Conv.	39.8%	48.1%	12.1%	0.67
			Ideas	40.4%	48.0%	11.6%	0.76
9	68219	69,710	Conv.	36.0%	46.7%	17.3%	0.50
			Ideas	41.9%	44.8%	13.3%	0.67
10	69030	54,817	Conv.	27.4%	42.3%	30.3%	0.38
			Ideas	32.9%	45.6%	21.6%	0.62

Note: For ECR items, target performance for QWK is a value greater than 0.70. There are no target performance metrics for exact agreement. N HSAS is the number of human-scored responses.

With respect to SMD values, many ECR items in the low confidence sample also tend to fall below target performance (Table 29).

Table 29. Performance of ASE with respect to SMD on ECR items in the low confidence sample

Grade	Item ID	N HSAS	Dim.	Max Score	Mean		SD		SMD	
					HS	AS	HS	AS	H1H2	HSAS
3	12624	53,564	Conv.	4	1.94	1.78	1.20	1.08	-0.03	0.14
			Ideas	6	2.37	2.62	1.20	1.17	-0.02	-0.21
4	12628	45,565	Conv.	4	1.50	1.59	1.14	0.90	0.00	-0.08
			Ideas	6	2.09	2.52	1.43	1.08	0.00	-0.33
5	12647	66,749	Conv.	4	1.87	2.09	1.38	1.06	-0.00	-0.18

Grade	Item ID	N HSAS	Dim.	Max Score	Mean		SD		SMD	
					HS	AS	HS	AS	H1H2	HSAS
			Ideas	6	2.55	2.75	1.78	1.44	-0.00	-0.12
6	12674	69,199	Conv.	4	1.11	1.03	1.29	0.99	0.00	0.06
			Ideas	6	1.89	2.06	1.94	1.41	-0.00	-0.10
7	61507	96,136	Conv.	4	2.44	2.50	1.24	1.05	0.01	-0.05
			Ideas	6	3.21	3.20	1.44	1.31	0.01	0.01
8	73974	75,873	Conv.	4	2.26	2.03	1.36	1.10	0.01	0.19
			Ideas	6	2.67	2.71	1.61	1.35	0.00	-0.02
9	68219	69,710	Conv.	4	2.27	1.83	1.22	0.89	0.00	0.41
			Ideas	6	2.67	2.78	1.47	1.14	-0.00	-0.08
10	69030	54,817	Conv.	4	2.02	1.85	1.46	1.03	-0.00	0.14
			Ideas	6	2.18	2.12	1.61	1.31	-0.00	0.04

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses.

Table 30 and 31 presents the dimension-level agreement statistics on the rubric scale for both human raters and the two models. Table 30 presents the Exact Agreement and Quadratic Weighted Kappa (QWK) of human-human agreement (H1H2), human-machine agreement (H1M1 and H2M2), and the difference between the two, for each ECR item. The average H1H2 exact agreement across items and dimensions was 12% lower in the low confidence sample relative to the random sample, and the average H1H2 QWK agreement was .27 lower. The H1M1 and H2M2 agreements showed a slightly larger drop (14% and 19% for M1H1 and M2H2 exact agreement, .31 and .30 for QWK agreement).

Table 30. Performance of ASE compared to human-human agreement, with respect to Exact Agreement and QWK, on the ECR rubric level scores in the low confidence sample

				H1H2	H1M1		H2M2		H1H2	H1M1		H2M2	
Grade	Item ID	N	Dim.	EA	EA	diff	EA	diff	QWK	QWK	diff	QWK	diff
3	12624	53,564	Conv.	58.6%	54.3%	-4.4%	62.8%	4.1%	0.48	0.45	-0.03	0.56	0.08
3	12624	53,564	Ideas	62.8%	64.7%	1.9%	63.2%	0.4%	0.53	0.54	0.02	0.56	0.03
4	12628	45,565	Conv.	55.2%	53.2%	-2.0%	55.2%	-0.1%	0.33	0.28	-0.06	0.35	0.02
4	12628	45,565	Ideas	51.4%	50.1%	-1.2%	51.4%	0.1%	0.42	0.36	-0.06	0.41	-0.01
5	12647	66,749	Conv.	54.1%	46.5%	-7.6%	53.5%	-0.6%	0.41	0.36	-0.05	0.48	0.06
5	12647	66,749	Ideas	51.6%	46.6%	-5.0%	53.7%	2.1%	0.54	0.53	-0.01	0.56	0.02
6	12674	69,199	Conv.	61.6%	57.4%	-4.2%	59.8%	-1.8%	0.43	0.36	-0.08	0.40	-0.04
6	12674	69,199	Ideas	60.0%	50.0%	-10.0%	48.6%	-11.4%	0.64	0.56	-0.08	0.53	-0.12
7	61507	96,136	Conv.	53.5%	53.8%	0.3%	57.8%	4.4%	0.41	0.44	0.03	0.47	0.06
7	61507	96,136	Ideas	53.5%	56.8%	3.3%	56.7%	3.2%	0.53	0.57	0.04	0.58	0.05
8	73974	75,873	Conv.	56.0%	54.6%	-1.3%	57.2%	1.2%	0.51	0.52	0.01	0.54	0.03

				H1H2	H1M1		H2M2		H1H2	H1M1		H2M2	
Grade	Item ID	N	Dim.	EA	EA	diff	EA	diff	QWK	QWK	diff	QWK	diff
8	73974	75,873	Ideas	52.7%	55.9%	3.2%	58.0%	5.3%	0.59	0.63	0.04	0.63	0.04
9	68219	69,710	Conv.	55.7%	49.4%	-6.3%	56.4%	0.6%	0.40	0.29	-0.11	0.42	0.02
9	68219	69,710	Ideas	53.6%	55.7%	2.1%	57.3%	3.6%	0.51	0.54	0.02	0.52	0.01
10	69030	54,817	Conv.	48.2%	40.7%	-7.5%	41.7%	-6.5%	0.32	0.27	-0.05	0.25	-0.07
10	69030	54,817	Ideas	49.2%	47.6%	-1.6%	49.7%	0.5%	0.46	0.46	0.00	0.52	0.06

Note: Target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10. N HSAS is the number of human-scored responses. H1M1 reflects the model 1 performance relative to rater 1. H2M2 refers to model 2 performance relative to rater 2

Table 31 displays the performance of ASE for the low confidence sample, with respect to the SMD statistic. Note here that the two human raters tend to assign similar scores, on average, but that the two models assign different scores, on average for many items and dimensions.

Table 31. Performance of ASE compared to human-human agreement, with respect to SMD, on the ECR rubric level scores in the low confidence sample

				Mean				SD				SMD		
Grade	Item ID	N	Dim.	H1	H2	M1	M2	H1	H2	M1	M2	H1H2	H1M1	H2M2
3	12624	53,564	Conv.	0.97	0.99	0.75	1.01	0.69	0.68	0.65	0.67	-0.03	0.34	-0.03
3	12624	53,564	Ideas	1.19	1.20	1.19	1.43	0.68	0.68	0.64	0.66	-0.02	-0.00	-0.34
4	12628	45,565	Conv.	0.75	0.75	0.81	0.79	0.66	0.66	0.53	0.62	0.00	-0.10	-0.06
4	12628	45,565	Ideas	1.05	1.05	1.29	1.24	0.80	0.80	0.67	0.63	0.00	-0.32	-0.27
5	12647	66,749	Conv.	0.95	0.95	0.94	1.13	0.76	0.76	0.75	0.66	-0.00	0.01	-0.25
5	12647	66,749	Ideas	1.29	1.29	1.32	1.40	0.95	0.95	0.84	0.82	-0.00	-0.03	-0.12
6	12674	69,199	Conv.	0.57	0.57	0.48	0.46	0.71	0.71	0.61	0.60	0.00	0.13	0.17
6	12674	69,199	Ideas	0.96	0.96	1.00	1.05	1.01	1.00	0.84	0.73	-0.00	-0.04	-0.10
7	61507	96,136	Conv.	1.23	1.22	1.11	1.44	0.70	0.71	0.68	0.61	0.01	0.17	-0.32
7	61507	96,136	Ideas	1.61	1.60	1.51	1.69	0.80	0.80	0.72	0.75	0.01	0.13	-0.12
8	73974	75,873	Conv.	1.14	1.14	1.04	0.99	0.76	0.76	0.68	0.65	0.01	0.14	0.21
8	73974	75,873	Ideas	1.35	1.35	1.30	1.39	0.88	0.88	0.78	0.72	0.00	0.06	-0.05
9	68219	69,710	Conv.	1.14	1.14	0.87	0.95	0.70	0.70	0.49	0.58	0.00	0.44	0.30
9	68219	69,710	Ideas	1.34	1.34	1.25	1.53	0.82	0.82	0.69	0.64	-0.00	0.12	-0.26
10	69030	54,817	Conv.	1.00	1.00	0.82	1.06	0.82	0.82	0.66	0.52	-0.00	0.23	-0.10
10	69030	54,817	Ideas	1.07	1.07	1.07	1.05	0.88	0.88	0.68	0.80	-0.00	0.00	0.03

Note: Target performance for SMD is within +/- 0.15. N HSAS is the number of human-scored responses.

All Responses

As described earlier, 71.8% of final scores were generated by ASE, while 28.2% were scored by human raters. Excluding condition codes, we can compare the final scores on all responses to engine and human scores on random routed sample. We should expect these to be similar because the random sample is intended to reflect the entire population of students for each item.

Table 32 presents descriptive statistics of SCR items, presenting means and standard deviations of all scored responses, alongside those of the Random Percent sample. The "All Scored" scores are based upon the hybrid scoring process. The "Rand. HS" reflects the score assigned by the human raters on the random percent sample. The "Rand. AS" reflects the score assigned by ASE. In general, means across both samples (and for both human and engine scores in the Random Percent sample) are similar. Score point distributions for all responses may be found in Appendix E.

Table 32. Descriptive statistics of all scored SCR responses compared to the random percent sample, by item

Grade	Item ID	Max Score	N		Mean			SD		
			Total	Random Percent	All Scored	Rand. HS	Rand. AS	All Scored	Rand. HS	Rand. AS
3	114749	1	334,998	33,514	0.44	0.45	0.45	0.50	0.50	0.50
3	83640	2	342,123	34,273	0.87	0.81	0.91	0.75	0.75	0.76
4	114768	1	351,149	35,110	0.35	0.35	0.35	0.48	0.48	0.48
4	91650	2	361,289	36,482	1.04	1.02	1.03	0.75	0.75	0.76
5	114786	1	363,926	36,724	0.59	0.57	0.58	0.49	0.49	0.49
5	84308	2	369,562	37,194	0.75	0.74	0.75	0.79	0.78	0.78
6	114807	1	382,183	38,393	0.65	0.65	0.65	0.48	0.48	0.48
6	2224	2	386,569	38,821	1.25	1.22	1.27	0.79	0.79	0.79
7	114822	1	386,328	38,537	0.41	0.41	0.41	0.49	0.49	0.49
7	90459	2	391,312	39,190	1.26	1.22	1.22	0.77	0.77	0.77
8	114840	1	391,888	39,303	0.65	0.65	0.67	0.48	0.48	0.47
8	89173	2	393,912	39,478	1.20	1.18	1.19	0.74	0.75	0.74
9	113231	1	459,235	45,861	0.45	0.45	0.45	0.50	0.50	0.50
9	90632	2	472,957	47,355	1.37	1.36	1.39	0.71	0.73	0.70
10	113258	1	442,559	44,213	0.58	0.58	0.59	0.49	0.49	0.49
10	89405	2	447,453	45,037	1.32	1.30	1.27	0.73	0.74	0.73

Note: HS and AS refer to human and automated scores, respectively, in the Random Percent sample.

Table 33 presents descriptive statistics of ECR item dimensions across all scored responses, alongside those of the random percent sample. While we see more variation in the mean scores than with SCR items, the standard deviations are also relatively larger.

Table 33. Descriptive statistics of all scored ECR responses compared to the random percent sample, by item

				N		Mean			SD		
Grade	Item ID	Dim.	Max Score	Total	Random Percent	All Scored	Rand. HS	Rand. AS	All Scored	Rand. HS	Rand. AS
3	12624	Conv.	4	293,004	29,625	1.45	1.54	1.39	1.41	1.49	1.37
3	12624	Ideas	6	293,004	29,625	1.94	1.93	1.99	1.58	1.59	1.59
4	12628	Conv.	4	317,242	31,955	1.87	1.79	1.88	1.52	1.49	1.49
4	12628	Ideas	6	317,242	31,955	2.56	2.51	2.63	1.91	1.92	1.86
5	12647	Conv.	4	338,165	34,128	1.28	1.27	1.33	1.53	1.54	1.47
5	12647	Ideas	6	338,165	34,128	1.70	1.73	1.75	1.93	1.99	1.85
6	12674	Conv.	4	356,226	35,671	1.61	1.64	1.59	1.60	1.59	1.56
6	12674	Ideas	6	356,226	35,671	2.48	2.57	2.52	2.22	2.26	2.12
7	61507	Conv.	4	364,844	36,722	2.00	1.93	2.03	1.55	1.51	1.51
7	61507	Ideas	6	364,844	36,722	2.64	2.57	2.66	1.98	1.90	1.96
8	73974	Conv.	4	363,928	36,407	2.12	2.16	2.06	1.56	1.57	1.51
8	73974	Ideas	6	363,928	36,407	2.68	2.68	2.70	1.96	1.99	1.91
9	68219	Conv.	4	416,421	42,086	1.89	1.92	1.80	1.61	1.62	1.56
9	68219	Ideas	6	416,421	42,086	2.54	2.47	2.56	2.13	2.12	2.09
10	69030	Conv.	4	409,947	41,233	2.32	2.22	2.31	1.66	1.63	1.62
10	69030	Ideas	6	409,947	41,233	2.90	2.80	2.91	2.10	2.04	2.08

Note: HS and AS refer to human and automated scores, respectively, in the Random Percent sample.

Conclusion and Next Steps

Overall, the results suggest that the hybrid scoring design is providing accurate, reliable, and fair scoring. All items met our full set of performance criteria on the full random sample.

Routing for both low confidence and condition code routing are performing adequately. The low confidence routing performances indicate that the engine is not performing well on these responses, which suggests that the confidence model and threshold is identifying responses that are difficult to score and should be routed for human scoring. The condition code routing agreements indicate that responses scored with the Out of Vocabulary condition code show very high agreements with the human raters. The other two condition codes performed adequately but will continue to be refined to improve agreements with the human raters.

Areas of future consideration include research into further refining the overall hybrid scoring design. This includes ensuring that hand-scores are returned quickly enough to reprogram the engine earlier in the test window. It also includes examining the impact of not using the original model for routing low confidence or condition code responses, and instead reserving that routing only for the final reprogrammed model. In order to ensure that 25% of responses are routed under this approach we can examine whether to increase the threshold for low confidence routing or increase the percentage of responses in the random percent routed sample. We will also examine changing the Unusual Score condition code to allow for these responses to be routed to the typical human rater pool, rather than the expert rater pool. The Out of Vocabulary condition code could potentially be considered for non-routing.

References

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

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.

Appendices

Appendix A: Score Point Distributions on the Operational Held-out Validation Sample

Table A1. Comparison of score distributions (in percentage) in SCR items generated by human raters and ASE in the held-out validation sample

Grade	Item ID	N	Rater	0	1	2
3	114749	1363	Human	55	45	
			Auto	56	44	
3	83640	1008	Human	38	41	21
			Auto	36	41	24
4	114768	1401	Human	64	36	
			Auto	65	35	
4	91650	1024	Human	31	40	29
			Auto	29	40	32
5	114786	1280	Human	45	55	
			Auto	44	56	
5	84308	800	Human	55	27	18
			Auto	52	31	16
6	114807	1774	Human	36	64	
			Auto	36	64	
6	2224	1439	Human	24	31	46
			Auto	23	27	50
7	114822	1782	Human	60	40	
			Auto	60	40	
7	90459	1541	Human	25	35	40
			Auto	24	35	41
8	114840	1955	Human	33	67	
			Auto	33	67	
8	89173	1933	Human	22	40	39
			Auto	21	40	39
9	113231	3462	Human	55	45	
			Auto	55	45	
9	90632	2161	Human	14	32	54
			Auto	13	33	54

Grade	Item ID	N	Rater	0	1	2
10	113258	2260	Human	38	62	
			Auto	40	60	
10	89405	1919	Human	17	39	44
			Auto	18	34	48

Note: Values represent percentages.

Table A2. Comparison of score distributions (in percentage) in ECR items generated by human raters and ASE in the held-out validation sample

Grade	Item ID	N	Dim.	Rater	0	1	2	3	4	5	6
3	12624	847	Conv.	Human	42	16	17	14	11		
				Auto	42	17	20	14	7		
3	12624	847	Ideas	Human	28	13	30	12	11	5	2
				Auto	26	12	32	14	12	3	1
4	12628	1207	Conv.	Human	29	14	20	17	20		
				Auto	28	19	16	19	18		
4	12628	1207	Ideas	Human	23	12	15	15	15	12	8
				Auto	18	15	16	15	18	10	7
5	12647	1186	Conv.	Human	48	14	10	12	15		
				Auto	46	15	14	11	13		
5	12647	1186	Ideas	Human	41	13	11	11	10	9	5
				Auto	40	17	12	9	11	7	3
6	12674	834	Conv.	Human	40	13	14	16	17		
				Auto	45	12	12	19	13		
6	12674	834	Ideas	Human	31	11	12	11	12	12	11
				Auto	31	15	11	10	15	7	11
7	61507	785	Conv.	Human	34	14	16	17	19		
				Auto	33	13	15	20	19		
7	61507	785	Ideas	Human	26	12	21	12	10	10	8
				Auto	27	12	20	12	11	9	9
8	73974	1056	Conv.	Human	26	12	17	18	27		
				Auto	27	14	17	18	25		
8	73974	1056	Ideas	Human	22	12	17	13	15	11	10
				Auto	20	13	18	12	16	12	9
9	68219	1194	Conv.	Human	32	8	17	16	27		
				Auto	32	10	14	20	23		
9	68219	1194	Ideas	Human	29	6	12	11	16	13	13

Grade	Item ID	N	Dim.	Rater	0	1	2	3	4	5	6
				Auto	25	9	12	11	20	11	12
10	69030	376	Conv.	Human	24	11	11	19	35		
				Auto	24	9	11	17	40		
10	69030	376	Ideas	Human	20	7	11	16	19	14	14
				Auto	19	10	10	15	15	15	16

Note: Values represent percentages.

Appendix B: Score Point Distributions on the Random Sample

Table B1. Comparison of score distributions (in percentage) in SCR items generated by human raters and ASE in the random percent sample

Grade	Item ID	N	Rater	0	1	2
3	114749	33514	Human	55	45	
			Auto	55	45	
3	83640	34273	Human	39	40	21
			Auto	34	41	25
4	114768	35110	Human	65	35	
			Auto	65	35	
4	91650	36482	Human	27	44	29
			Auto	27	43	30
5	114786	36724	Human	43	57	
			Auto	42	58	
5	84308	37194	Human	47	32	21
			Auto	46	33	21
6	114807	38393	Human	35	65	
			Auto	35	65	
6	2224	38821	Human	22	33	45
			Auto	21	30	49
7	114822	38537	Human	59	41	
			Auto	59	41	
7	90459	39190	Human	21	36	43
			Auto	21	35	43
8	114840	39303	Human	35	65	
			Auto	33	67	
8	89173	39478	Human	21	40	39
			Auto	20	41	39
9	113231	45861	Human	55	45	
			Auto	55	45	
9	90632	47355	Human	15	34	51
			Auto	13	35	52
10	113258	44213	Human	42	58	
			Auto	41	59	
10	89405	45037	Human	17	36	47
			Auto	17	38	44

Grade	Item ID	N	Rater	0	1	2

Note: Values represent percentages.

Table B2. Comparison of score distributions (in percentage) in ECR items generated by human raters and ASE in the random percent sample

Grade	Item ID	N	Dim.	Rater	0	1	2	3	4	5	6
3	12624	29625	Conv.	Human	38	15	18	13	16		
				Auto	38	19	18	16	9		
3	12624	29625	Ideas	Human	27	12	28	13	13	4	2
				Auto	26	11	31	13	13	4	2
4	12628	31955	Conv.	Human	30	13	22	16	19		
				Auto	27	17	17	19	20		
4	12628	31955	Ideas	Human	24	11	16	14	18	10	7
				Auto	18	14	15	15	20	12	6
5	12647	34128	Conv.	Human	52	10	12	10	15		
				Auto	45	16	13	13	13		
5	12647	34128	Ideas	Human	47	9	10	9	13	7	5
				Auto	38	18	13	10	11	7	4
6	12674	35671	Conv.	Human	40	12	14	15	20		
				Auto	41	12	13	18	17		
6	12674	35671	Ideas	Human	33	9	9	9	14	12	15
				Auto	27	13	12	11	15	9	12
7	61507	36722	Conv.	Human	28	12	19	19	21		
				Auto	25	13	17	21	23		
7	61507	36722	Ideas	Human	21	10	21	13	17	10	8
				Auto	20	11	20	13	15	11	10
8	73974	36407	Conv.	Human	25	10	17	17	30		
				Auto	24	14	18	19	25		
8	73974	36407	Ideas	Human	22	10	17	13	16	12	10
				Auto	19	11	20	13	18	12	9
9	68219	42086	Conv.	Human	34	8	16	15	26		
				Auto	34	11	15	20	20		
9	68219	42086	Ideas	Human	32	7	13	10	17	11	10
				Auto	28	10	11	11	20	10	10
10	69030	41233	Conv.	Human	27	9	13	17	34		
				Auto	25	8	13	17	36		
10	69030	41233	Ideas	Human	23	8	13	12	19	13	11

Grade	Item ID	N	Dim.	Rater	0	1	2	3	4	5	6
				Auto	21	9	13	12	18	13	14

Note: Values represent percentages.

Appendix C: Student Group Performance on the Random Sample for Each Item

Table C1. Student group performance of on SCR items with respect to Exact Agreement, QWK, and SMD in the random percent sample, disaggregated by student group

Grade	Item ID	Max Score	Std. Group	Exact Agreement			QWK			SMD	
				H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
3	114749	1	Female	96.4%	97.2%	0.8%	0.93	0.94	0.02	0.00	-0.01
			Male	96.2%	97.1%	0.9%	0.92	0.94	0.02	-0.02	-0.01
			Black	96.5%	97.2%	0.8%	0.91	0.94	0.02	-0.01	-0.00
			Latino	96.4%	97.2%	0.7%	0.92	0.94	0.02	-0.02	-0.01
			White	96.2%	97.0%	0.9%	0.92	0.94	0.02	0.00	-0.01
			Low SES	96.7%	97.3%	0.7%	0.92	0.94	0.02	-0.01	-0.01
			EB	96.8%	97.2%	0.4%	0.93	0.94	0.01	-0.01	-0.01
3	83640	2	Female	77.7%	79.1%	1.3%	0.79	0.81	0.02	0.01	-0.14
			Male	78.2%	80.2%	2.0%	0.79	0.81	0.03	-0.00	-0.11
			Black	80.0%	81.8%	1.8%	0.81	0.82	0.01	0.01	-0.11
			Latino	78.4%	79.9%	1.6%	0.78	0.81	0.03	-0.00	-0.12
			White	76.4%	77.9%	1.5%	0.79	0.79	0.00	0.02	-0.15
			Low SES	79.0%	80.6%	1.6%	0.78	0.81	0.03	0.01	-0.12
			EB	78.4%	80.2%	1.9%	0.78	0.81	0.03	0.00	-0.13
4	114768	1	Female	96.4%	97.1%	0.7%	0.92	0.94	0.02	0.00	0.01
			Male	95.7%	97.4%	1.6%	0.90	0.94	0.04	0.01	0.01
			Black	97.1%	97.6%	0.5%	0.92	0.94	0.02	-0.01	0.01
			Latino	95.8%	97.2%	1.3%	0.90	0.93	0.03	0.01	0.01
			White	96.1%	97.3%	1.2%	0.92	0.95	0.02	-0.00	0.01
			Low SES	96.1%	97.3%	1.1%	0.90	0.93	0.03	0.01	0.01
			EB	95.1%	97.2%	2.1%	0.88	0.93	0.05	0.00	0.02
4	91650	2	Female	66.5%	72.4%	5.9%	0.67	0.74	0.07	0.01	-0.05
			Male	69.2%	72.5%	3.2%	0.69	0.74	0.05	-0.00	0.00
			Black	68.3%	74.1%	5.8%	0.67	0.75	0.08	0.03	-0.02
			Latino	67.8%	72.7%	4.9%	0.68	0.74	0.07	0.01	-0.02
			White	67.1%	70.6%	3.5%	0.68	0.73	0.05	-0.01	-0.01
			Low SES	68.1%	73.3%	5.3%	0.67	0.75	0.08	0.02	-0.01
			EB	67.6%	73.3%	5.7%	0.67	0.75	0.07	0.02	-0.03
5	114786	1	Female	91.5%	93.5%	2.0%	0.83	0.87	0.04	-0.01	-0.02
			Male	90.9%	93.4%	2.5%	0.82	0.86	0.05	0.01	-0.01
			Black	92.5%	93.4%	0.9%	0.85	0.87	0.02	-0.03	-0.00
			Latino	91.2%	93.4%	2.2%	0.82	0.87	0.04	0.01	-0.02

Grade	Item ID	Max Score	Std. Group	Exact Agreement			QWK			SMD	
				H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
			White	90.7%	93.3%	2.6%	0.80	0.85	0.06	0.00	-0.02
			Low SES	91.4%	93.3%	2.0%	0.83	0.87	0.04	-0.00	-0.01
			EB	92.0%	92.9%	0.9%	0.84	0.86	0.02	-0.01	-0.02
5	84308	2	Female	70.9%	74.3%	3.4%	0.73	0.77	0.05	0.00	-0.02
			Male	72.1%	75.7%	3.6%	0.73	0.77	0.04	0.01	0.01
			Black	71.8%	76.8%	5.1%	0.68	0.76	0.07	-0.02	-0.02
			Latino	72.9%	75.8%	3.0%	0.72	0.76	0.04	0.01	0.01
			White	68.8%	72.9%	4.1%	0.71	0.77	0.06	0.00	-0.01
			Low SES	72.5%	76.5%	4.0%	0.70	0.75	0.06	0.00	0.00
			EB	73.7%	76.3%	2.6%	0.71	0.75	0.04	0.03	0.01
6	114807	1	Female	91.5%	94.6%	3.1%	0.81	0.87	0.07	-0.01	0.01
			Male	91.6%	94.4%	2.8%	0.83	0.88	0.06	0.00	-0.00
			Black	90.4%	94.2%	3.8%	0.81	0.88	0.08	-0.02	0.01
			Latino	91.3%	94.5%	3.2%	0.82	0.88	0.07	-0.00	-0.00
			White	92.4%	94.4%	2.0%	0.81	0.86	0.05	-0.01	0.00
			Low SES	91.0%	94.3%	3.3%	0.81	0.88	0.07	0.00	0.00
			EB	91.7%	94.6%	2.9%	0.83	0.89	0.06	-0.01	0.00
6	2224	2	Female	75.0%	81.7%	6.7%	0.76	0.83	0.06	0.02	-0.08
			Male	76.2%	81.1%	4.9%	0.80	0.84	0.05	0.01	-0.04
			Black	73.9%	80.3%	6.4%	0.77	0.84	0.07	0.03	-0.04
			Latino	74.4%	80.4%	5.9%	0.78	0.83	0.05	0.01	-0.06
			White	77.8%	82.5%	4.8%	0.78	0.83	0.05	0.02	-0.07
			Low SES	73.9%	80.0%	6.1%	0.77	0.83	0.06	0.02	-0.05
			EB	74.0%	79.8%	5.8%	0.79	0.84	0.05	0.01	-0.05
7	114822	1	Female	96.8%	97.8%	1.0%	0.93	0.96	0.02	-0.01	-0.00
			Male	96.8%	98.0%	1.1%	0.93	0.96	0.02	0.00	-0.01
			Black	97.5%	98.1%	0.6%	0.94	0.96	0.01	0.00	-0.00
			Latino	96.7%	97.9%	1.2%	0.93	0.95	0.03	-0.00	-0.01
			White	96.6%	97.6%	1.0%	0.93	0.95	0.02	0.00	-0.01
			Low SES	96.5%	97.9%	1.3%	0.92	0.95	0.03	-0.00	-0.00
			EB	97.0%	98.0%	1.0%	0.92	0.95	0.03	0.00	-0.00
7	90459	2	Female	73.7%	77.4%	3.7%	0.74	0.78	0.04	0.01	-0.02
			Male	73.8%	77.8%	4.0%	0.76	0.81	0.04	-0.00	0.01
			Black	73.7%	77.3%	3.6%	0.75	0.80	0.05	0.01	0.01
			Latino	73.4%	77.3%	3.9%	0.76	0.80	0.04	0.00	0.01
			White	73.5%	77.5%	4.0%	0.72	0.77	0.04	0.02	-0.02
			Low SES	73.7%	76.8%	3.1%	0.76	0.80	0.04	0.02	0.01
			EB	74.2%	77.8%	3.6%	0.76	0.81	0.05	-0.00	0.01

Grade	Item ID	Max Score	Std. Group	Exact Agreement			QWK			SMD	
				H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
8	114840	1	Female	90.7%	93.9%	3.3%	0.78	0.85	0.07	-0.03	-0.05
			Male	88.9%	93.2%	4.3%	0.77	0.86	0.09	-0.00	-0.04
			Black	89.8%	92.5%	2.8%	0.79	0.84	0.06	-0.04	-0.03
			Latino	89.2%	93.5%	4.3%	0.78	0.86	0.09	-0.01	-0.03
			White	90.2%	93.7%	3.5%	0.75	0.83	0.08	-0.00	-0.06
			Low SES	89.0%	93.2%	4.2%	0.77	0.86	0.08	-0.01	-0.03
			EB	87.4%	93.4%	5.9%	0.75	0.87	0.12	-0.01	-0.02
8	89173	2	Female	74.9%	79.9%	5.1%	0.74	0.80	0.05	0.01	-0.03
			Male	77.1%	81.3%	4.3%	0.79	0.83	0.05	0.01	-0.00
			Black	76.5%	80.3%	3.8%	0.77	0.82	0.05	-0.02	-0.00
			Latino	75.9%	81.2%	5.3%	0.77	0.83	0.05	0.01	-0.01
			White	75.3%	79.3%	4.0%	0.77	0.80	0.04	0.01	-0.03
			Low SES	76.4%	81.2%	4.8%	0.78	0.83	0.05	0.00	-0.01
			EB	75.9%	81.6%	5.7%	0.77	0.83	0.06	0.00	-0.01
9	113231	1	Female	97.0%	98.2%	1.2%	0.94	0.96	0.02	0.00	0.01
			Male	97.4%	98.2%	0.7%	0.95	0.96	0.02	-0.01	0.01
			Black	97.3%	97.9%	0.6%	0.94	0.96	0.01	0.01	0.01
			Latino	97.4%	98.3%	0.8%	0.95	0.96	0.02	-0.01	0.01
			White	96.7%	98.1%	1.3%	0.93	0.96	0.03	-0.00	0.00
			Low SES	97.4%	98.1%	0.8%	0.94	0.96	0.02	-0.01	0.01
			EB	97.7%	98.3%	0.6%	0.95	0.96	0.02	0.00	0.01
9	90632	2	Female	79.8%	82.9%	3.0%	0.77	0.80	0.03	-0.00	-0.06
			Male	78.8%	81.3%	2.5%	0.81	0.83	0.02	0.00	-0.04
			Black	76.4%	81.0%	4.6%	0.78	0.82	0.05	-0.01	-0.05
			Latino	78.6%	81.2%	2.6%	0.80	0.82	0.02	-0.00	-0.04
			White	81.2%	83.5%	2.3%	0.76	0.79	0.02	0.01	-0.05
			Low SES	77.8%	80.8%	3.0%	0.80	0.82	0.02	-0.00	-0.05
			EB	77.8%	80.3%	2.5%	0.81	0.82	0.01	-0.01	-0.03
10	113258	1	Female	92.6%	93.7%	1.2%	0.84	0.87	0.02	0.02	-0.02
			Male	91.5%	93.5%	2.0%	0.83	0.87	0.04	0.03	-0.03
			Black	91.7%	92.5%	0.8%	0.83	0.85	0.02	0.03	-0.03
			Latino	91.5%	93.4%	1.9%	0.83	0.87	0.04	0.03	-0.03
			White	92.9%	94.3%	1.4%	0.85	0.87	0.03	0.02	-0.02
			Low SES	91.4%	93.1%	1.7%	0.83	0.86	0.03	0.03	-0.02
			EB	90.8%	93.2%	2.4%	0.81	0.86	0.05	0.04	-0.02
10	89405	2	Female	83.9%	80.4%	-3.5%	0.83	0.79	-0.04	-0.01	0.03
			Male	84.2%	78.6%	-5.5%	0.86	0.81	-0.05	-0.02	0.05
			Black	83.1%	78.4%	-4.7%	0.85	0.81	-0.04	-0.05	0.05

Grade	Item ID	Max Score	Std. Group	Exact Agreement			QWK			SMD	
				H1H2	HSAS	Diff.	H1H2	HSAS	Diff.	H1H2	HSAS
			Latino	83.3%	78.6%	-4.8%	0.85	0.81	-0.04	-0.02	0.05
			White	85.3%	80.7%	-4.6%	0.84	0.79	-0.05	0.00	0.03
			Low SES	83.5%	78.6%	-4.9%	0.85	0.81	-0.04	-0.02	0.06
			EB	83.6%	77.8%	-5.8%	0.85	0.80	-0.05	-0.01	0.06

Table C2. Student group performance of on ECR items with respect to Exact Agreement, QWK, and SMD in the random percent sample, disaggregated by student group

					Agreement			QWK	SMD	
Grade	Item ID	Dim.	Max Score	Std. Group	HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS	H1H2	HSAS
3	12624	Conv.	4	Female	55.9%	36.6%	7.4%	0.84	-0.01	0.10
				Male	58.8%	33.6%	7.6%	0.82	-0.00	0.11
				Black	61.4%	32.8%	5.9%	0.84	-0.01	0.09
				Latino	59.5%	33.4%	7.0%	0.83	-0.00	0.08
				White	53.9%	37.9%	8.3%	0.82	-0.02	0.15
				Low SES	60.9%	32.4%	6.7%	0.83	-0.01	0.08
				EB	58.2%	34.6%	7.3%	0.82	0.00	0.08
3	12624	Ideas	6	Female	58.2%	34.5%	7.3%	0.87	-0.01	-0.04
				Male	59.7%	33.3%	7.0%	0.86	0.01	-0.02
				Black	62.6%	31.0%	6.4%	0.87	0.02	-0.04
				Latino	59.6%	33.1%	7.3%	0.86	-0.00	-0.04
				White	57.5%	35.3%	7.1%	0.86	-0.01	-0.01
				Low SES	61.5%	31.7%	6.9%	0.86	-0.00	-0.04
				EB	60.2%	32.4%	7.5%	0.85	0.02	-0.05
4	12628	Conv.	4	Female	53.4%	37.6%	9.0%	0.82	-0.01	-0.06
				Male	55.5%	36.2%	8.3%	0.82	0.00	-0.06
				Black	58.5%	33.6%	7.9%	0.83	0.00	-0.03
				Latino	54.6%	36.7%	8.7%	0.82	-0.01	-0.07
				White	52.1%	38.9%	9.0%	0.81	0.00	-0.05
				Low SES	55.9%	35.5%	8.7%	0.81	-0.00	-0.06
				EB	55.0%	36.3%	8.7%	0.82	0.01	-0.09
4	12628	Ideas	6	Female	48.2%	40.9%	10.9%	0.86	0.00	-0.07
				Male	50.4%	39.3%	10.3%	0.87	0.00	-0.06
				Black	51.9%	37.8%	10.3%	0.86	-0.00	-0.06
				Latino	49.8%	39.4%	10.8%	0.86	0.00	-0.08
				White	48.1%	41.4%	10.6%	0.86	0.01	-0.05
				Low SES	50.7%	38.7%	10.6%	0.86	0.00	-0.08
				EB	50.0%	39.0%	11.0%	0.86	0.01	-0.09

					Agreement			QWK	SMD	
Grade	Item ID	Dim.	Max Score	Std. Group	HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS	H1H2	HSAS
5	12647	Conv.	4	Female	58.4%	31.3%	10.3%	0.83	0.00	-0.05
				Male	62.4%	28.6%	9.0%	0.83	0.01	-0.04
				Black	65.8%	26.1%	8.1%	0.83	0.01	-0.03
				Latino	62.2%	28.5%	9.3%	0.82	0.01	-0.05
				White	56.5%	33.0%	10.5%	0.82	0.01	-0.03
				Low SES	64.3%	27.0%	8.7%	0.82	0.01	-0.04
				EB	63.0%	28.3%	8.7%	0.82	0.01	-0.05
5	12647	Ideas	6	Female	52.0%	35.7%	12.3%	0.86	0.01	-0.02
				Male	55.5%	33.6%	10.9%	0.86	0.01	-0.00
				Black	59.8%	30.6%	9.6%	0.86	0.01	0.01
				Latino	55.7%	33.5%	10.9%	0.86	0.01	-0.01
				White	49.7%	37.3%	13.0%	0.85	0.01	-0.02
				Low SES	58.6%	31.3%	10.1%	0.86	0.01	-0.00
				EB	57.0%	32.1%	11.0%	0.85	0.02	-0.01
6	12674	Conv.	4	Female	53.7%	34.1%	12.3%	0.80	-0.01	0.03
				Male	58.6%	30.7%	10.7%	0.82	-0.00	0.04
				Black	58.8%	30.2%	11.0%	0.80	-0.01	0.06
				Latino	57.5%	31.0%	11.5%	0.80	-0.01	0.03
				White	52.3%	35.3%	12.4%	0.80	-0.01	0.04
				Low SES	59.0%	30.1%	10.9%	0.80	-0.01	0.04
				EB	60.4%	28.6%	11.0%	0.79	-0.01	0.03
6	12674	Ideas	6	Female	49.9%	36.0%	14.1%	0.87	-0.00	0.02
				Male	54.9%	33.4%	11.7%	0.89	0.01	0.03
				Black	54.6%	33.0%	12.4%	0.88	-0.00	0.02
				Latino	53.7%	33.4%	12.9%	0.88	0.00	0.03
				White	49.0%	37.2%	13.8%	0.87	-0.00	0.03
				Low SES	54.3%	33.1%	12.6%	0.88	0.00	0.02
				EB	56.0%	31.6%	12.4%	0.88	-0.00	0.02
7	61507	Conv.	4	Female	53.7%	37.1%	9.2%	0.82	-0.00	-0.09
				Male	57.6%	33.9%	8.4%	0.84	0.02	-0.06
				Black	56.9%	34.3%	8.7%	0.83	0.02	-0.07
				Latino	56.2%	35.1%	8.8%	0.83	0.01	-0.08
				White	53.1%	37.7%	9.1%	0.81	0.00	-0.05
				Low SES	56.9%	34.4%	8.7%	0.83	0.01	-0.07
				EB	58.5%	32.7%	8.8%	0.82	0.01	-0.10
7	61507	Ideas	6	Female	51.5%	40.0%	8.6%	0.89	0.00	-0.06
				Male	56.2%	35.9%	7.9%	0.90	0.01	-0.04

					Agreement			QWK	SMD	
Grade	**Item ID**	**Dim.**	**Max Score**	**Std. Group**	**HSAS Exact**	**HSAS Adj.**	**HSAS Non-adj.**	**HSAS**	**H1H2**	**HSAS**
				Black	57.0%	35.0%	8.0%	0.90	-0.00	-0.05
				Latino	55.2%	36.9%	8.0%	0.89	0.01	-0.05
				White	50.3%	40.9%	8.8%	0.88	0.01	-0.03
				Low SES	56.2%	36.1%	7.8%	0.89	0.01	-0.05
				EB	57.4%	35.5%	7.1%	0.89	0.00	-0.06
8	73974	Conv.	4	Female	57.6%	34.9%	7.5%	0.85	-0.01	0.07
				Male	60.7%	32.9%	6.4%	0.87	0.00	0.05
				Black	58.8%	33.9%	7.3%	0.85	0.00	0.06
				Latino	58.6%	34.3%	7.1%	0.86	-0.00	0.05
				White	58.0%	34.7%	7.4%	0.85	-0.01	0.10
				Low SES	59.1%	33.8%	7.1%	0.86	-0.00	0.06
				EB	59.7%	33.5%	6.7%	0.85	0.00	0.04
8	73974	Ideas	6	Female	51.3%	40.1%	8.6%	0.89	-0.00	-0.00
				Male	56.1%	36.7%	7.2%	0.91	0.00	-0.02
				Black	55.4%	37.2%	7.4%	0.89	-0.00	-0.03
				Latino	54.9%	37.2%	7.9%	0.89	-0.00	-0.02
				White	50.5%	41.0%	8.4%	0.89	-0.00	0.02
				Low SES	55.5%	36.9%	7.6%	0.89	-0.00	-0.02
				EB	57.6%	34.5%	7.8%	0.89	-0.00	-0.03
9	68219	Conv.	4	Female	59.2%	33.8%	7.0%	0.87	-0.00	0.08
				Male	62.9%	30.1%	7.0%	0.87	0.00	0.08
				Black	62.4%	30.7%	6.9%	0.87	-0.01	0.08
				Latino	62.3%	30.8%	6.9%	0.87	0.00	0.07
				White	56.9%	35.2%	7.9%	0.85	-0.01	0.13
				Low SES	62.8%	30.2%	7.0%	0.87	0.00	0.07
				EB	67.2%	26.7%	6.1%	0.86	0.02	0.04
9	68219	Ideas	6	Female	56.6%	35.8%	7.6%	0.92	0.00	-0.05
				Male	61.3%	32.3%	6.4%	0.92	0.00	-0.04
				Black	61.4%	31.9%	6.7%	0.92	-0.00	-0.05
				Latino	60.7%	32.5%	6.8%	0.92	0.01	-0.05
				White	54.1%	38.2%	7.7%	0.91	-0.00	-0.04
				Low SES	61.6%	31.8%	6.6%	0.92	0.00	-0.05
				EB	66.3%	27.4%	6.3%	0.92	0.00	-0.05
10	69030	Conv.	4	Female	58.3%	30.4%	11.4%	0.81	-0.02	-0.07
				Male	59.3%	29.0%	11.7%	0.83	0.00	-0.05
				Black	56.5%	30.8%	12.7%	0.80	-0.00	-0.05
				Latino	57.4%	30.5%	12.1%	0.82	-0.01	-0.05

					Agreement			QWK	SMD	
Grade	Item ID	Dim.	Max Score	Std. Group	HSAS Exact	HSAS Adj.	HSAS Non-adj.	HSAS	H1H2	HSAS
				White	60.1%	29.3%	10.6%	0.79	-0.01	-0.06
				Low SES	57.3%	30.5%	12.2%	0.81	-0.00	-0.05
				EB	57.9%	29.1%	12.9%	0.78	0.00	-0.06
10	69030	Ideas	6	Female	48.0%	39.5%	12.5%	0.86	-0.00	-0.05
				Male	51.6%	36.5%	11.9%	0.88	0.01	-0.05
				Black	51.5%	36.3%	12.2%	0.86	0.01	-0.03
				Latino	50.5%	37.1%	12.4%	0.87	0.00	-0.05
				White	46.8%	41.0%	12.3%	0.86	-0.00	-0.06
				Low SES	51.2%	36.5%	12.4%	0.87	-0.00	-0.04
				EB	54.0%	33.9%	12.0%	0.85	0.00	-0.05

Appendix D: Score Point Distributions on the Low Confidence Sample

Table D1. Comparison of score distributions (in percentage) in SCR items generated by human raters and ASE in the low confidence sample

Grade	Item ID	N	Rater	0	1	2
3	114749	43402	Human	47	53	
			Auto	43	57	
3	83640	63833	Human	14	57	29
			Auto	8	49	42
4	114768	54393	Human	66	34	
			Auto	67	33	
4	91650	63919	Human	19	54	26
			Auto	21	50	29
5	114786	68051	Human	54	46	
			Auto	58	42	
5	84308	71441	Human	32	49	19
			Auto	26	59	15
6	114807	63103	Human	70	30	
			Auto	69	31	
6	2224	83019	Human	20	39	42
			Auto	14	39	47
7	114822	59890	Human	60	40	
			Auto	59	41	
7	90459	69562	Human	11	49	40
			Auto	18	57	25
8	114840	63580	Human	60	40	
			Auto	51	49	
8	89173	62126	Human	20	54	26
			Auto	20	56	24
9	113231	75321	Human	62	38	
			Auto	62	38	
9	90632	80524	Human	17	51	32
			Auto	12	50	38
10	113258	62885	Human	64	36	
			Auto	55	45	
10	89405	71885	Human	10	38	51
			Auto	14	55	31

Grade	Item ID	N	Rater	0	1	2

Note: Values represent percentages.

Table D2. Comparison of score distributions (in percentage) in ECR items generated by human raters and ASE in the low confidence sample

Grade	Item ID	N	Dim.	Rater	0	1	2	3	4	5	6
3	12624	53564	Conv.	Human	15	18	35	20	11		
				Auto	13	29	28	26	4		
3	12624	53564	Ideas	Human	9	8	40	23	17	2	0
				Auto	6	6	37	26	21	3	0
4	12628	45565	Conv.	Human	25	23	35	13	5		
				Auto	12	34	39	15	1		
4	12628	45565	Ideas	Human	18	16	27	20	16	3	1
				Auto	4	13	32	34	15	2	0
5	12647	66749	Conv.	Human	24	14	27	20	15		
				Auto	7	22	36	26	10		
5	12647	66749	Ideas	Human	20	9	18	18	21	8	5
				Auto	5	15	28	23	17	9	4
6	12674	69199	Conv.	Human	48	17	18	11	7		
				Auto	37	33	21	9	1		
6	12674	69199	Ideas	Human	41	10	11	12	16	6	5
				Auto	17	17	30	21	10	5	1
7	61507	96136	Conv.	Human	9	13	26	28	24		
				Auto	4	14	27	38	17		
7	61507	96136	Ideas	Human	5	7	21	22	29	12	5
				Auto	1	9	21	27	25	14	3
8	73974	75873	Conv.	Human	15	14	24	24	23		
				Auto	11	21	31	31	7		
8	73974	75873	Ideas	Human	12	12	22	19	22	9	3
				Auto	3	18	24	26	19	10	1
9	68219	69710	Conv.	Human	12	10	33	26	18		
				Auto	9	23	47	21	1		
9	68219	69710	Ideas	Human	11	8	25	23	24	7	2
				Auto	4	8	25	35	24	3	0
10	69030	54817	Conv.	Human	23	15	20	20	22		
				Auto	11	24	40	21	5		
10	69030	54817	Ideas	Human	20	14	29	15	14	5	3
				Auto	6	26	41	14	4	8	1

Grade	Item ID	N	Dim.	Rater	0	1	2	3	4	5	6

Note: Values represent percentages.

Appendix E: Score Point Distributions across All Scored Responses

Table E1. Descriptive statistics of final scores across all scored SCR responses, by item

Grade	Item ID	N	0	1	2
3	114749	334,998	56	44	
3	83640	342,123	35	42	22
4	114768	351,149	65	35	
4	91650	361,289	26	44	30
5	114786	363,926	41	59	
5	84308	369,562	47	31	22
6	114807	382,183	35	65	
6	2224	386,569	22	31	47
7	114822	386,328	59	41	
7	90459	391,312	20	34	46
8	114840	391,888	35	65	
8	89173	393,912	20	41	40
9	113231	459,235	55	45	
9	90632	472,957	14	35	51
10	113258	442,559	42	58	
10	89405	447,453	16	36	48

Table E2. Descriptive statistics of final scores across all scored ECR responses, by item

Grade	Item ID	N	Dim.	0	1	2	3	4	5	6
3	12624	293,004	Conv.	38	17	19	15	12		
3	12624	293,004	Ideas	26	11	31	13	13	4	2
4	12628	317,242	Conv.	29	15	17	18	21		
4	12628	317,242	Ideas	21	14	14	12	20	12	6
5	12647	338,165	Conv.	51	12	11	11	15		
5	12647	338,165	Ideas	44	14	10	9	12	7	5
6	12674	356,226	Conv.	43	9	12	18	18		
6	12674	356,226	Ideas	32	11	8	9	16	10	13
7	61507	364,844	Conv.	27	13	17	17	25		
7	61507	364,844	Ideas	21	11	20	11	17	10	11
8	73974	363,928	Conv.	25	12	17	17	29		
8	73974	363,928	Ideas	21	10	19	11	18	11	9
9	68219	416,421	Conv.	35	8	13	20	23		
9	68219	416,421	Ideas	30	9	11	9	20	11	11

Grade	Item ID	N	Dim.	0	1	2	3	4	5	6
10	69030	409,947	Conv.	27	7	11	16	39		
10	69030	409,947	Ideas	23	7	11	12	19	13	14