CRASE5® for ACT Writing Technical Report

Scott W. Wood, Sungjin Nam, and Dongmei Li

I. Introduction

Starting in October 2022, ACT began using its automated scoring engine, CRASE®, to provide one of the two rater scores for ACT writing essays in the ACT International program. Since then, CRASE has been used for other programs, including ACT District (Spring 2023), ACT State (Fall 2023), and ACT National Digital (December 2023).

To support this decision, ACT created a significant research agenda. Researchers conducted multiple proof-of-concept studies to evaluate the accuracy of CRASE scores compared to those of human scorers. The results were published in June 2023 as the research report *CRASE+® for ACT Writing Technical Report*.

Since CRASE began to score ACT writing, there has been a desire to add new engine functionalities in order to expand the kinds of essays CRASE can handle. The new functionalities include automatically detecting off-topic essays, automatically detecting disturbing content, leveraging modern model-fitting approaches, and providing information about the confidence of the engine in assigning scores. A new version of CRASE, called CRASE5, was developed in 2025 to incorporate these new functionalities.

The primary purpose of this report is to replicate the studies listed in the CRASE+ technical report using the new generic scoring models produced in CRASE5. The results presented in this report, especially when compared to the corresponding results in the original report, should provide validity evidence that the updated CRASE5 generic scoring models perform as well as the original CRASE+ generic scoring models.

This report follows the organization of the CRASE+ technical report. The next section contains a brief overview of CRASE and CRASE5. Section III discusses the data and processes used to train and validate the engine, while Section IV provides validation results. Section V contains a subgroup analysis of the CRASE5 scores. Section VI describes how condition codes are handled by CRASE5.

II. Background: Automated Scoring and CRASE5

Automated scoring (or automated essay scoring) is the use of a computer algorithm to emulate hand scoring behavior on constructed-response or essay items. The scoring algorithm is called the engine. Preparing the scoring algorithm for operational use is called training the engine. There are four parts to a scoring engine: a means of reading text data, a preprocessor that standardizes and initially processes the text, a means of extracting the quantitative characteristics of the text (called features), and a means of mapping these characteristics to hand scoring data. For advanced engines using neural networks, the feature extraction and the mapping of features to hand scoring data may be combined into one process.

CRASE, short for Constructed Response Automated Scoring Engine, was created in 2007 for a U.S. state's summative assessment program. The system has since been enhanced to include scoring methods for additional types of free-response items and to incorporate new technologies in text processing and modeling. CRASE has been used operationally in multiple state testing programs (formative and summative) and in many research programs, including a U.S. Department of Education Enhanced Assessment Grant. As mentioned in the introduction, CRASE has been used in the scoring queue for selected online ACT writing tests since 2022.

When CRASE was first used on ACT writing, it was called CRASE+ and represented the fourth major version of the CRASE software. With the construction of a new engine that gives the engine trainer greater flexibility in developing scoring models, ACT proposed the use of CRASE5 starting in September 2025. All analyses presented in this report were produced using scores from CRASE5.

This report assumes a basic familiarity with automated scoring concepts. For readers new to automated scoring, the CRASE research team recommends the following resources:

Lottridge, S., Burkhardt, A., & Boyer, M. (2020). Digital module 18: Automated scoring. *Educational Measurement: Issues and Practice*, *39*(3), 141–142. https://doi.org/10.1111/emip.12388

McCaffrey, D., Casablanca, J., Ricker-Pedley, K., Lawless, R., & Wendler, C. (2021). *Best practices for constructed-response scoring*. ETS. https://www.ets.org/content/dam/ets-org/pdfs/about/cr_best_practices.pdf

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions.* Routledge.

Wood, S., Yao, E., Haisfield, L., & Lottridge, S. (2021). *Establishing standards of best practice in automated scoring*. ACT. https://www.act.org/content/dam/act/unsecured/documents/R2100-auto-scoringstandards-2021-07.pdf

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press.

III. Methods for Engine Training and Validation

Data

Data from hand-scored essays are required to train the CRASE engine. These data should be collected under authentic testing conditions, if possible, and must be representative of the population of examinees expected to submit essays in the future.

This report uses the same ACT writing data from the original training and validation studies conducted prior to 2023. Readers can find details about the data, the training sample, and the blind-validation sample in the *CRASE+® for ACT Writing Technical Report*. This section will summarize key information about the data.

The training and blind-validation essays came from three sources: the September 2020 ACT International administration, the October 2020 ACT International administration, and selected Spring 2021 State and District administrations. Approximately two thirds of the records came from the State and District administrations. Only essays obtained via online administrations were included. Essays assigned condition codes by hand scorers were not eligible for the training or blind-validation samples. In all, approximately 14,000 essays with valid hand scores were available for training and validation, with less than 1% of the essays being excluded due to condition codes.

Information about hand scoring score point distributions, examinee gender, examinee Hispanic status, and examinee race/ethnicity can be found in the CRASE+ technical report on pages 5 and 6.

Training and Validation Samples

Recall that a generic scoring model is an automated scoring model built using essay data from multiple writing prompts with the goal of using the model on essay data from comparable writing prompts. (The alternative is a prompt-specific model, where the model is built using essay data from a single writing prompt with the goal of using the model on essay data from only that prompt.) Generic scoring models allow for consistent scoring regardless of prompt. Generic models also allow prompts added to the prompt bank to be automatically scored. Plus, prompts in the prompt bank that have few field-testing results may still use automated scoring via generic models. Generic scoring models have been used for ACT writing since 2022.

The training sample is used to determine the model of best fit. The blind-validation sample, being blind to the model-training process, is a new set of data used to evaluate the model of best fit by replicating how the model would perform in operational practice on new essays.

To ensure that the blind-validation sample contained essays from prompts not seen by the engine during training, the CRASE research team allocated data to the training and blind-validation samples by prompt. Rules were established to ensure that International prompts and State and District prompts were evenly represented in each sample. The training sample

contained 8,862 essays across 16 writing prompts. The blind-validation sample contained 5,128 essays across 11 prompts.

Further information about the gender, Hispanic status, and race/ethnicity distributions for the training and blind-validation samples can be found in the CRASE+ technical report.

Engine Training

The research team trained the CRASE5 engine in the same way it trained CRASE+. The default set of 39 writing features was computed for each essay in the training sample, and these features were mapped to the hand scoring raw score of record.

For ACT writing, the score of record is defined as follows:

- If Rater 1 and Rater 2 assign the same score to an essay for a given domain, the final score of record is the sum of the two raters' scores.
- If Rater 1 and Rater 2 assign scores that are within one point of each other (for example, a 3 and a 4), then the final score of record is the sum of the two raters' scores.
- If Rater 1 and Rater 2 assign scores that differ by more than 1 point (for example, a 2 and a 5), then a third rater is assigned to perform a resolution read. The final raw score is provided by the resolution reader.
- In all cases, an examinee can earn a score of record between 2 and 12, inclusive, on each writing domain.

After producing the scoring models, the research team resubmitted the training essays to the model to produce predictions. Because gradient-boosted regression models were used for training, the returned predictions are decimals between approximately 1.000 and 6.000 (for example, 3.58943820). To convert these undiscretized predictions to discretized scores, a discretization rule is required.

In the original generic models, discretization cut scores were created using the score point distribution for Rater 1. For the 1–6 scale, the 1–2 cut score was defined so that the percentage of examinees in the training sample getting a 1 from Rater 1 (A%) would be the same as the lowest A% of the predicted scores. The 2–3 cut score was defined so that the percentage of examinees in the training sample getting a 2 from Rater 1 (B%) would be the same as the next lowest percentage (B%) of the predicted scores.

CRASE5 can use multiple sets of discretization cut scores. For the 2–12 scale, the above process was modified by using the percentage of examinees getting a score of record of 2, the percentage of examinees getting a score of record of 3, etc. and then setting the discretization cut scores to copy that distribution using 2–12 scores.

For this project, CRASE5 results included both the discretized predicted scores on the 1–6 scale and the discretized predicted scores on the 2–12 scale. Note that the original report considered scores only on the 1–6 scale. The 2–12 scale is included in this report for further proposed research.

Engine Evaluation

Automated scoring models are evaluated using distributional metrics and agreement metrics from the blind-validation sample. Distributional metrics include the score point distribution, mean, and standard deviation of the scores produced by Rater 1, Rater 2, and CRASE5. The expectation is that the CRASE5 distributional metrics will be similar to those produced by Rater 1 and Rater 2.

Distributional metrics include the standardized mean difference, or SMD. This metric is defined as the mean score from Rater 1 minus the mean score from Rater 2, divided by the pooled standard deviation. If the absolute value of the SMD is less than or equal to 0.15, then the means of the two distributions are similar enough to be used in practice (Williamson et al., 2012).

Agreement statistics are used to evaluate rater reliability (that is, the degree of agreement between two independent raters). The exact agreement rate is the percentage of essays to which two raters have assigned the same score. The adjacent agreement rate is the percentage of essays to which the two raters have assigned scores that are different but within 1 point of each other. When raters are scoring on a 1–6 scale, ACT standards require that the exact agreement rate be 60% or higher and the sum of the exact and adjacent agreement rates be 95% or higher.

Automated scoring professionals will report kappa and quadratic weighted kappa (QWK) in evaluations. These metrics, similar to correlations, are measures of rater agreement that account for the fact that raters sometimes agree simply by chance. Both metrics incorporate penalties for disagreements, but QWK incorporates greater penalties when the raters differ by larger amounts. Industry standards recommend that the human–computer QWK be greater than or equal to 0.70 in order for models to be used operationally (Williamson et al., 2012).

Later sections of this report will cover other forms of model evaluation, such as analysis for subgroup differences.

IV. Results for Engine Training and Validation

Baseline Results on the 1–6 Scale

When researchers are developing a new version of an automated scoring engine, they hope that the scoring models from the new version will be comparable to those from the old engine. CRASE5 was designed to produce the same kinds of features and models in the same way as CRASE+. However, subtle changes in third-party libraries and CRASE functions will cause some minimal differences in the scores assigned by both engines.

To confirm that the CRASE5 models function similarly to those from CRASE+, the research team produced automated scores on the 1–6 scale using the new engine. These scores were compared with the original hand scores from Rater 1 and Rater 2. We expect that the CRASE5 score point distribution will be similar to the distributions from Rater 1 and Rater 2. Further, we expect the R1–CRASE5 and R2–CRASE5 agreement statistics to match or exceed the R1–R2 agreement statistics, as was the case using CRASE+.

Tables 1a and 1b contain the distributional metrics and agreement metrics for the first writing domain, Ideas and Analysis. As shown in Table 1a, the CRASE5 score point distribution is very similar to the Rater 1 and Rater 2 distributions. Further, the mean and standard deviation are consistent with the mean and standard deviation from the Rater 1 and Rater 2 distributions.

Table 1b shows that the R1–CRASE5 exact agreement rate (71.1%) and the R2–CRASE5 exact agreement rate (71.6%) exceed both the R1–R2 exact agreement rate (68.0%) and ACT's minimum exact agreement threshold of 60%. The exact-plus-adjacent agreement rates for all pairs of raters exceed 99%, which easily surpasses ACT's minimum threshold of 95%. The R1–CRASE5 and R2–CRASE5 quadratic weighted kappas are both 0.85, which slightly exceeds the R1–R2 QWK of 0.83. In practice, QWKs greater than 0.70 are expected for operational use.

Table 1a. Distributional Metrics, Generic Model, Domain 1 (Ideas and Analysis)

Score	Rater 1	Rater 2	CRASE5
Mean	3.5	3.5	3.4
SD	1.0	1.0	1.0
1	3.6%	3.5%	3.3%
2	11.3%	11.3%	12.3%
3	32.3%	32.9%	33.8%
4	39.0%	38.7%	38.7%
5	12.6%	12.1%	10.6%
6	1.3%	1.6%	1.3%

Note. N = 5,128

Table 1b. Agreement Metrics, Generic Model, Domain 1 (Ideas and Analysis)

Metric	Rater 1–Rater 2	Rater 1–CRASE5	Rater 2–CRASE5
|SMD|	0.00	0.05	0.05
***SD* ratio**	1.00	1.02	1.02
Exact agreement	68.0%	71.1%	71.6%
Adjacent agreement	31.3%	29.5%	27.8%
Nonadjacent agreement	0.6%	0.4%	0.6%
Kappa	.55	.59	.60
QWK	.83	.85	.85
Correlation	.83	.85	.85

Note. N = 5,128

Tables 2a and 2b contain the distributional metrics and agreement metrics for the second writing domain, Development and Style. For all three raters, the mean Development and Style score is one to two tenths of a point lower than the mean Ideas and Analysis score. As before, the CRASE5 score point distribution, mean, and standard deviation are all similar to the respective statistics for Rater 1 and Rater 2.

Table 2b shows that the R1–CRASE5 exact agreement rate (72.0%) and the R2–CRASE5 exact agreement rate (72.6%) exceed both the R1–R2 exact agreement rate (68.4%) and ACT's minimum exact agreement threshold of 60%. The exact-plus-adjacent agreement rates for all pairs of raters exceed 99%. The R1–CRASE5 and R2–CRASE5 quadratic weighted kappas are both 0.85, which slightly exceeds the R1–R2 QWK of 0.83.

Table 2a. Distributional Metrics, Generic Model, Domain 2 (Development and Style)

Score	Rater 1	Rater 2	CRASE5
Mean	3.3	3.3	3.3
SD	1.0	1.0	1.0
1	3.9%	3.8%	3.7%
2	17.6%	17.6%	17.4%
3	36.0%	35.9%	38.0%
4	33.0%	33.2%	32.5%
5	9.0%	8.7%	7.6%
6	0.5%	0.7%	0.7%

Note. N = 5,128

Table 2b. Agreement Metrics, Generic Model, Domain 2 (Development and Style)

Metric	Rater 1–Rater 2	Rater 1–CRASE5	Rater 2–CRASE5
|SMD|	0.00	0.02	0.02
***SD* ratio**	1.00	1.02	1.02
Exact agreement	68.4%	72.0%	72.6%
Adjacent agreement	31.0%	27.6%	26.9%
Nonadjacent agreement	0.6%	0.4%	0.5%
Kappa	.56	.61	.62
QWK	.83	.85	.85
Correlation	.83	.85	.85

Note. N = 5,128

Tables 3a and 3b contain the distributional metrics and agreement metrics for the third writing domain, Organization. As before, the CRASE5 score point distribution, mean, and standard deviation are all similar to the respective statistics for Rater 1 and Rater 2.

Table 3b shows that the R1–CRASE5 exact agreement rate (71.6%) and the R2–CRASE5 exact agreement rate (71.7%) exceed both the R1–R2 exact agreement rate (68.4%) and ACT's minimum exact agreement threshold. The exact-plus-adjacent agreement rates for all pairs of raters exceed 99.5%. The R1–CRASE5 and R2–CRASE5 quadratic weighted kappas are both 0.84, which slightly exceeds the R1–R2 QWK of 0.83.

Table 3a. Distributional Metrics, Generic Model, Domain 3 (Organization)

Score	Rater 1	Rater 2	CRASE5
Mean	3.4	3.4	3.4
SD	1.0	1.0	1.0
1	3.6%	3.4%	3.3%
2	12.3%	12.4%	13.5%
3	33.7%	34.0%	34.9%
4	39.0%	38.7%	38.4%
5	10.6%	10.5%	9.0%
6	0.8%	0.9%	0.8%

Note. N = 5,128

Table 3b. Agreement Metrics, Generic Model, Domain 3 (Organization)

Metric	Rater 1–Rater 2	Rater 1–CRASE5	Rater 2–CRASE5
|SMD|	0.00	0.04	0.04
***SD* ratio**	1.00	1.02	1.01
Exact agreement	68.4%	71.6%	71.7%
Adjacent agreement	31.1%	27.9%	27.8%
Nonadjacent agreement	0.5%	0.4%	0.4%
Kappa	.55	.60	.60
QWK	.83	.84	.84
Correlation	.83	.85	.85

Note. N = 5,128

Lastly, Tables 4a and 4b contain the distributional metrics and agreement metrics for the final writing domain, Language Use and Conventions. The mean Language Use and Conventions score for each of the three raters is one to two tenths of a point higher than that for the other domains. As before, the CRASE5 score point distribution, mean, and standard deviation are all similar to the respective statistics for Rater 1 and Rater 2.

Table 4b shows that the R1–CRASE5 exact agreement rate (69.6%) and the R2–CRASE5 exact agreement rate (70.2%) exceed both the R1–R2 exact agreement rate (68.1%) and ACT's minimum exact agreement threshold. The exact-plus-adjacent agreement rates for all pairs of raters exceed 99%. The R1–CRASE5 and R2–CRASE5 quadratic weighted kappas are both 0.82, which slightly exceeds the R1–R2 QWK of 0.81.

Table 4a. Distributional Metrics, Generic Model, Domain 4 (Language Use and Conventions)

Score	Rater 1	Rater 2	CRASE5
Mean	3.6	3.6	3.6
SD	0.9	0.9	0.9
1	2.0%	1.9%	1.9%
2	8.2%	8.0%	9.0%
3	31.9%	32.2%	33.5%
4	43.2%	43.3%	42.4%
5	13.3%	12.9%	11.9%
6	1.5%	1.8%	1.4%

Note. N = 5,128

Table 4b. Agreement Metrics, Generic Model, Domain 4 (Language Use and Conventions)

Metric	Rater 1–Rater 2	Rater 1–CRASE5	Rater 2–CRASE5
|SMD|	0.01	0.04	0.06
***SD* ratio**	1.00	1.01	1.01
Exact agreement	68.1%	69.6%	70.2%
Adjacent agreement	31.2%	30.0%	29.3%
Nonadjacent agreement	0.7%	0.4%	0.5%
Kappa	.54	.56	.57
QWK	.81	.82	.82
Correlation	.81	.82	.82

Note. N = 5,128

By all accounts, the CRASE5 models designed to score essays on a 1–6 scale are comparable to the Rater 1 and Rater 2 scores. Further, the R1–CRASE5 and R2–CRASE5 agreement rates are comparable to or better than their respective R1–R2 agreement rates. Therefore, we conclude that the CRASE5 1–6 models are appropriate for operational use.

Overall, the four generic models performed according to ACT and automated scoring standards. Because different prompts in the blind-validation sample go to different populations

(International or State and District), it is important to review the agreement metrics of the blind-validation sample by prompt. Tables 5a–5k contain the R1–CRASE5 standardized mean differences, exact agreement rates, and QWKs by writing domain and writing prompt. Metrics that do not meet ACT and automated scoring thresholds are marked with an asterisk.

All but one QWK exceeded the .70 threshold (prompt I114_01019, Domain 4). All but one exact agreement rate exceeded the 60% threshold (prompt I114_01071, Domain 3). Ten of the forty-four SMDs are outside the range of −0.15 to 0.15, affecting one or more domains on 3 of the 11 prompts. These results are an improvement over the CRASE+ models, which exhibited four exact agreement rates less than 60%, thirteen SMDs outside the desired range, and 5 prompts affected by large SMDs.

During operational use of the generic models, if it is determined that the R1–CRASE5 agreement rates are not meeting ACT and automated scoring thresholds for certain prompts, ACT can have these prompts scored by at least two hand scorers. This will ensure that examinees receive the best quality scoring, regardless of prompt.

Table 5a. R1–CRASE5 Agreement Metrics for I114_00617, by Domain (*n* = 185)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.27*	61.6%	.80
Development and Support	0.26*	71.4%	.83
Organization	0.24*	67.6%	.81
Language Use and Conventions	0.22*	67.0%	.80

*Does not meet ACT and automated scoring thresholds

Table 5b. R1–CRASE5 Agreement Metrics for I114_00789, by Domain (*n* = 1,447)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.00	71.9%	.85
Development and Support	0.04	71.9%	.85
Organization	0.00	72.7%	.85
Language Use and Conventions	0.01	69.8%	.81

Table 5c. R1–CRASE5 Agreement Metrics for I114_00794, by Domain (*n* = 69)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.08	72.5%	.81
Development and Support	0.03	79.7%	.86
Organization	0.08	75.4%	.83
Language Use and Conventions	0.04	71.0%	.76

Table 5d. R1–CRASE5 Agreement Metrics for I114_00915, by Domain (*n* = 1,618)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.15	69.0%	.82
Development and Support	0.12	70.8%	.82
Organization	0.13	70.5%	.82
Language Use and Conventions	0.14	68.9%	.79

Table 5e. R1–CRASE5 Agreement Metrics for I114_00934, by Domain (*n* = 662)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.03	71.3%	.81
Development and Support	0.05	72.4%	.82
Organization	0.01	71.8%	.80
Language Use and Conventions	0.02	69.3%	.77

Table 5f. R1–CRASE5 Agreement Metrics for I114_00939, by Domain (*n* = 414)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.03	78.0%	.89
Development and Support	0.01	77.5%	.88
Organization	0.03	75.6%	.86
Language Use and Conventions	0.05	76.3%	.86

Table 5g. R1–CRASE5 Agreement Metrics for I114_00972, by Domain (*n* = 211)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.09	75.4%	.87
Development and Support	0.06	77.3%	.89
Organization	0.09	73.9%	.86
Language Use and Conventions	0.13	73.5%	.84

Table 5h. R1–CRASE5 Agreement Metrics for I114_01019, by Domain (*n* = 158)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.17*	68.4%	.76
Development and Support	0.20*	70.3%	.76
Organization	0.18*	70.3%	.74
Language Use and Conventions	0.26*	60.1%	.65*

*Does not meet ACT and automated scoring thresholds

Table 5i. R1–CRASE5 Agreement Metrics for I114_01071, by Domain (*n* = 105)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.09	62.9%	.76
Development and Support	0.19*	61.9%	.77
Organization	0.17*	57.1%*	.72
Language Use and Conventions	0.14	60.0%	.74

*Does not meet ACT and automated scoring thresholds

Table 5j. R1–CRASE5 Agreement Metrics for I114_01075, by Domain (*n* = 124)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.06	76.6%	.87
Development and Support	0.04	68.5%	.83
Organization	0.05	71.0%	.83
Language Use and Conventions	0.05	75.0%	.83

Table 5k. R1–CRASE5 Agreement Metrics for I114_01152, by Domain (*n* = 135)

Writing domain	|SMD|	Exact	QWK
Ideas and Analysis	0.06	74.8%	.86
Development and Support	0.02	68.9%	.82
Organization	0.05	74.8%	.84
Language Use and Conventions	0.02	67.4%	.78

Baseline Results on the 2–12 Scale

To evaluate the CRASE5 2–12 models' distributional metrics, ACT compares the models' scores to the hand scoring score of record. Again, it is hoped that the score point distribution, mean, and standard deviation of CRASE5 2–12 scores are comparable to the respective statistics based on the scores of record.

Evaluating the agreement statistics is a bit more challenging. First, CRASE5 is attempting to predict a score on a scale with 11 possible score points, compared to the 6 possible score points of the 1–6 scale. As the number of score categories increases, exact agreement usually decreases, and QWK usually increases. A 60% exact agreement rate threshold may be appropriate for a 1–6 scale, but this may need to be adjusted for a 2–12 scale.

Second, we do not have a human–human agreement rate for the 2–12 scale to use for human–CRASE5 comparisons. There is only one 2–12 scale produced by hand scoring that can be used, which is the score of record. (While we could consider Rater 1 plus Rater 2 as a possible alternative, that would yield a variable so close to the score of record as to be effectively redundant. Likewise, resolution reads could be considered, but only a small number of records have a resolution read.)

Table 6a contains the distributional metrics for the first writing domain, Ideas and Analysis, on the 2–12 scale. The CRASE5 2–12 score point distributions, means, and standard deviations are comparable to those based on the scores of record.

Table 6b contains the score of record–CRASE5 (hereafter SR–CRASE5) agreement statistics. The exact agreement rate is 56.6%, while the exact-plus-adjacent agreement rate is 92.2%. The SR–CRASE5 QWK is 0.91. Since there is not an independent hand score on the 2–12 scale available for all blind-validation sample records, it is not possible to conclude that the SR–CRASE5 accuracy measures are comparable to results from human hand scorers using a 2–12 scale. Though a QWK above 0.90 is suitable for operational use, it would be up to stakeholders to determine whether an exact agreement of 56.6% and an exact-plus-adjacent agreement of 92.2% are acceptable given the size of the 2–12 scale.

Table 6a. Distributional Metrics, Generic Model, Domain 1 (Ideas and Analysis)

Score	Score of record	CRASE5
Mean	7.0	6.9
SD	1.9	1.9
2	2.7%	2.6%
3	1.4%	1.2%
4	7.8%	9.1%
5	5.1%	4.7%
6	24.5%	24.9%
7	11.3%	12.7%
8	29.7%	29.2%
9	7.5%	6.4%
10	7.8%	7.2%
11	1.5%	1.5%
12	0.6%	0.5%

Note. N = 5,128

Table 6b. Agreement Metrics, Generic Model, Domain 1 (Ideas and Analysis)

Metric	SR–CRASE5
|SMD|	0.03
***SD* ratio**	1.01
Exact agreement	56.6%
Adjacent agreement	35.5%
Nonadjacent agreement	7.8%
Kappa	.47
QWK	.91
Correlation	.91

Note. N = 5,128

Table 7a contains the distributional metrics for the second writing domain, Development and Style, on the 2–12 scale. The score point distributions, means, and standard deviations based on the score of record are comparable to those based on the CRASE5 2–12 scale.

Table 7b contains the SR–CRASE5 agreement statistics. The exact agreement rate is 58.5%, slightly higher than that for the first writing domain, while the exact-plus-adjacent agreement rate is 92.8%. The SR–CRASE5 QWK is 0.91. Once again, it would be up to stakeholders to determine whether an exact agreement of 58.5% and an exact-plus-adjacent agreement of 92.8% are acceptable given the size of the 2–12 scale.

Table 7a. Distributional Metrics, Generic Model, Domain 2 (Development and Style)

Score	Score of record	CRASE5
Mean	6.5	6.5
SD	1.9	1.9
2	3.1%	2.9%
3	1.5%	1.7%
4	13.5%	12.9%
5	6.6%	6.7%
6	27.0%	28.8%
7	11.8%	12.3%
8	24.2%	23.6%
9	6.1%	5.2%
10	5.3%	4.9%
11	0.7%	0.8%
12	0.2%	0.2%

Note. N = 5,128

Table 7b. Agreement Metrics, Generic Model, Domain 2 (Development and Style)

Metric	SR–CRASE5
|SMD|	0.02
***SD* ratio**	1.02
Exact agreement	58.5%
Adjacent agreement	34.3%
Nonadjacent agreement	7.2%
Kappa	.50
QWK	.91
Correlation	.91

Note. N = 5,128

Table 8a contains the distributional metrics for the third writing domain, Organization, on the 2–12 scale. The score point distributions, means, and standard deviations based on the score of record are comparable to those based on the CRASE5 2–12 scale.

Table 8b contains the SR–CRASE5 agreement statistics. The exact agreement rate is 57.5%, while the exact-plus-adjacent agreement rate is 92.5%. The SR–CRASE5 QWK is 0.90.

Table 8a. Distributional Metrics, Generic Model, Domain 3 (Organization)

Score	Score of record	CRASE5
Mean	6.9	6.8
SD	1.9	1.9
2	2.8%	2.7%
3	1.4%	1.4%
4	8.8%	9.2%
5	5.5%	5.1%
6	25.7%	26.5%
7	11.4%	13.1%
8	29.6%	28.9%
9	7.0%	6.1%
10	6.4%	5.6%
11	1.0%	1.0%
12	0.3%	0.3%

Note. N = 5,128

Table 8b. Agreement Metrics, Generic Model, Domain 3 (Organization)

Metric	SR–CRASE5
|SMD|	0.03
***SD* ratio**	1.02
Exact agreement	57.5%
Adjacent agreement	35.0%
Nonadjacent agreement	7.5%
Kappa	.48
QWK	.90
Correlation	.90

Note. N = 5,128

Table 9a contains the distributional metrics for the final writing domain, Language Use and Conventions, on the 2–12 scale. The score point distributions, means, and standard deviations based on the score of record are comparable to those based on the CRASE5 2–12 scale.

Table 9b contains the SR–CRASE5 agreement statistics. The exact agreement rate is 54.2%, the lowest of the four writing domain agreement rates, while the exact-plus-adjacent agreement rate is 90.9%. The SR–CRASE5 QWK is 0.88.

Table 9a. Distributional Metrics, Generic Model, Domain 4 (Language Use and Conventions)

Score	Score of record	CRASE5
Mean	7.3	7.2
SD	1.8	1.8
2	1.4%	1.3%
3	1.0%	0.9%
4	5.1%	6.2%
5	4.7%	4.3%
6	23.6%	24.2%
7	11.5%	13.3%
8	33.8%	32.4%
9	8.1%	7.4%
10	8.5%	7.8%
11	1.6%	1.5%
12	0.7%	0.7%

Note. N = 5,128

Table 9b. Agreement Metrics, Generic Model, Domain 4 (Language Use and Conventions)

Metric	SR–CRASE5
|SMD|	0.04
***SD* ratio**	1.01
Exact agreement	54.2%
Adjacent agreement	36.7%
Nonadjacent agreement	9.1%
Kappa	0.43
QWK	0.88
Correlation	0.88

Note. N = 5,128

If exact agreement rates above 54% on the 2–12 scale and exact-plus-adjacent agreement rates above 90.9% are considered acceptable by stakeholders, then there appears to be enough evidence that the 2–12 CRASE5 models can be used operationally. The CRASE5 score point distributions, mean, and standard deviation are very similar to those for the score of

record. Additionally, the QWKs for all four domains exceed 0.88, which would be considered well-fitting for a scale of this size.

V. Subgroup Analysis

Two standards in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014) directly promote subgroup analysis in automated scoring. Standard 3.8 applies to the scoring of all constructed response items, regardless of whether they are human or computer scored: "When tests require the scoring of constructed responses, test developers and/or users should collect and report evidence of the validity of score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores" (p. 66). The comment included with Standard 3.8 includes specific references to automated scoring: "Scoring algorithms need to be reviewed for potential sources of bias. The precision of scores and validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups of the intended population" (p. 67).

The other standard promoting subgroup analysis as an automated scoring practice is found in the comment included with Standard 4.19: "[Developers] may . . . collect independent judgments of the extent to which the resulting scores will accurately implement intended scoring rubrics and be free from bias for intended examinee subpopulations" (p. 92).

This section describes one common approach to subgroup analysis in automated scoring. Dubbed the ETS-style analysis, it is based on ETS's method of analyzing products like the GRE and the Praxis I, which use automated scoring (see, for example, Ramineni et al., 2012; Ramineni et al., 2015).

Methods

Using the 1–6 scale, we computed the following metrics for reported gender (male and female), reported Hispanic status (Hispanic and not Hispanic), and reported race/ethnicity (Asian, Black/African American, multiple, White, and blank/chose not to respond):

- The number of examinees in the subgroup
- The mean and standard deviation of the Rater 1 scores
- The mean and standard deviation of the Rater 2 scores
- The standardized mean difference between the Rater 1 and Rater 2 scores
- The (unweighted) kappa between the Rater 1 and Rater 2 scores
- The quadratic weighted kappa between the Rater 1 and Rater 2 scores
- The exact agreement rate between the Rater 1 and Rater 2 scores
- The sum of the Rater 1 and Rater 2 exact and adjacent agreement rates
- The Pearson correlation between the Rater 1 and Rater 2 scores
- The mean and standard deviation of the Rater 1 scores (repeated for convenience)
- The mean and standard deviation of the CRASE5 discretized scores
- The standardized mean difference between the Rater 1 and CRASE5 discretized scores

- The (unweighted) kappa between the Rater 1 and CRASE5 discretized scores
- The quadratic weighted kappa between the Rater 1 and CRASE5 discretized scores
- The exact agreement rate between the Rater 1 and CRASE5 discretized scores
- The sum of the Rater 1 and CRASE5 exact and adjacent agreement rates
- The Pearson correlation between the Rater 1 and CRASE5 discretized scores
- The mean and standard deviation of the Rater 1 scores (repeated again for convenience)
- The mean and standard deviation of the CRASE5 scores before they were discretized by the engine (in other words, the raw predictions from the regression model)
- The standardized mean difference between the Rater 1 scores and the CRASE5 undiscretized scores
- The Pearson correlation between the Rater 1 scores and the CRASE5 undiscretized scores
- The R1–CRASE5 (discretized) QWK minus the R1–R2 QWK
- The R1–CRASE5 (undiscretized) Pearson correlation minus the R1–R2 Pearson correlation

Values in the results will be marked with an asterisk if

- any SMD is larger than 0.10,
- any QWK is less than .70,
- any exact agreement rate is less than 60%,
- any exact-plus-adjacent agreement rate is less than 95%,
- any correlation is less than .70,
- the R1–CRASE5 QWK minus the R1–R2 QWK is less than −.10, or
- the R1–CRASE5 correlation minus the R1–R2 Pearson correlation is less than −.10.

These metrics and tests are equivalent to those appearing in various ETS automated scoring reports containing subgroup analyses.

Using the 2–12 scale and the scores of record from hand scoring, we computed the following metrics for reported gender (male and female), reported Hispanic status (Hispanic and not Hispanic), and reported race/ethnicity (Asian, Black/African American, multiple, White, and blank/chose not to respond):

- The number of examinees in the subgroup
- The mean and standard deviation of the scores of record
- The mean and standard deviation of the CRASE5 discretized scores
- The standardized mean difference between the scores of record and the CRASE5 discretized scores
- The (unweighted) kappa between the scores of record and the CRASE5 discretized scores
- The quadratic weighted kappa between the scores of record and the CRASE5 discretized scores

- The exact agreement rate between the scores of record and the CRASE5 discretized scores
- The sum of the exact and adjacent agreement rates for the scores of record and CRASE5
- The Pearson correlation between the scores of record and the CRASE5 discretized scores
- The mean and standard deviation of the scores of record (repeated for convenience)
- The mean and standard deviation of the CRASE5 scores before they were discretized by the engine (in other words, the raw predictions from the regression model) times 2
- The standardized mean difference between the scores of record and the CRASE5 undiscretized scores times 2
- The Pearson correlation between the scores of record and the CRASE5 undiscretized scores times 2

Values in the results based on the 2–12 scale will be marked with an asterisk if

- any SMD is larger than 0.10,
- any QWK is less than .70, or
- any correlation is less than .70.

Results

Tables 10–13 summarize the metrics for the 1–6 scale.

For all four domains, all gender subgroup metrics were within expectations. There do not appear to be any subgroup differences based on gender.

For all four domains, all Hispanic/non-Hispanic subgroup metrics were within expectations except for the Hispanic R1–CRASE5 (discretized) SMD for Domain 2 (0.20). Besides this one exception, there do not appear to be any subgroup differences based on Hispanic status.

Across the four domains, there was one instance when a race-based metric did not meet expectations: for Domain 2, the R1–CRASE5 (discretized) standardized mean difference for those identifying as multiple races was 0.14, which exceeded the 0.10 threshold. Besides this one exception, there do not appear to be any subgroup differences based on race.

Based on this subgroup analysis, subgroup differences are minimal when the 1–6 scale is used, and they do not significantly affect scoring accuracy. In fact, using CRASE5, we see only two metrics exceeding thresholds, compared to the four seen in the CRASE+ analysis.

Tables 14–17 summarize the metrics for the 2–12 scale.

For all four domains, all gender subgroup metrics were within expectations. There do not appear to be any subgroup differences based on gender.

For all four domains, all Hispanic/non-Hispanic subgroup metrics were within expectations except for the Hispanic R1–CRASE5 (discretized) SMD for Domain 2 (0.22). Besides this one exception, there do not appear to be any subgroup differences based on Hispanic status.

Across the four domains, there were two instances when a race-based metric did not meet expectations:

- For Domain 2, the R1–CRASE5 (discretized) standardized mean difference for those identifying as multiple races was 0.18, which exceeded the 0.10 threshold.
- For Domain 4, the R1–CRASE5 (undiscretized) standardized mean difference for those identifying as White was 0.11, which exceeded the 0.10 threshold.

Besides these two exceptions, there do not appear to be any subgroup differences based on race.

Based on this subgroup analysis, subgroup differences are minimal when the 2–12 scale is used, and they do not significantly affect scoring accuracy.

Table 10. ETS-Style Subgroup Analysis on Domain 1 Scores, Blind-Validation Sample

Group	*n*	H1 by H2									H1 by CRASE5 (discretized)									H1 by CRASE5 (unrounded)						Degradation	
		H1		H2		Stats					H1		CRASE5		Stats					H1		CRASE5		Stats		QWK	*r*
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	M	*SD*	M	*SD*	SMD	*r*	H1CRASE5 (rounded) – H1H2	H1CRASE5 (unrounded) – H1H2
All	5,128	3.5	1.0	3.5	1.0	0.00	.55	.83	68.0	99.4	.83	3.5	1.0	3.4	1.0	0.05	.59	.85	71.1	99.6	.85	3.5	1.0	3.5	0.9	0.02	.88	0.02	0.05
Male	2,440	3.4	1.0	3.4	1.0	0.01	.57	.85	68.8	99.3	.85	3.4	1.0	3.3	1.0	0.06	.60	.86	70.8	99.6	.86	3.4	1.0	3.4	0.9	0.03	.89	0.01	0.04
Female	2,580	3.6	1.0	3.6	1.0	0.01	.54	.82	67.7	99.4	.82	3.6	1.0	3.6	1.0	0.03	.59	.84	71.6	99.5	.84	3.6	1.0	3.6	0.8	0.00	.87	0.02	0.05
Hispanic	875	3.2	1.0	3.1	1.0	0.01	.58	.85	70.1	99.7	.85	3.2	1.0	3.1	1.0	0.01	.62	.85	73.1	99.2	.85	3.2	1.0	3.2	0.9	0.05	.88	0.00	0.03
Not Hisp.	4,028	3.6	1.0	3.6	1.0	0.00	.54	.82	67.6	99.3	.82	3.6	1.0	3.5	1.0	0.06	.59	.84	71.0	99.6	.84	3.6	1.0	3.5	0.9	0.04	.88	0.02	0.06
Blank	1,244	3.0	1.0	3.0	1.0	0.01	.55	.84	67.3	99.4	.84	3.0	1.0	2.9	1.0	0.04	.59	.84	70.5	99.2	.84	3.0	1.0	3.0	0.9	0.02	.88	0.00	0.04
Asian	583	3.9	1.0	3.9	1.0	0.02	.55	.82	68.3	99.1	.82	3.9	1.0	4.0	1.0	0.04	.53	.83	66.7	99.8	.83	3.9	1.0	3.9	0.9	0.02	.87	0.01	0.05
Black	346	3.1	1.0	3.0	1.0	0.04	.48	.78	63.9	98.6	.78	3.1	1.0	3.0	0.9	0.07	.58	.82	71.4	99.1	.83	3.1	1.0	3.1	0.8	0.01	.86	0.04	0.08
2+ Races	249	3.6	0.9	3.6	1.0	0.04	.54	.82	67.9	99.6	.82	3.6	0.9	3.5	1.0	0.09	.58	.85	70.7	100.0	.85	3.6	0.9	3.6	0.9	0.07	.90	0.03	0.08
White	1,878	3.8	0.9	3.8	0.9	0.00	.52	.77	68.2	99.4	.77	3.8	0.9	3.7	0.8	0.09	.57	.80	72.2	99.7	.80	3.8	0.9	3.7	0.7	0.08	.85	0.03	0.08

H1 = human rater 1, H2 = human rater 2, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 11. ETS-Style Subgroup Analysis on Domain 2 Scores, Blind-Validation Sample

Group	*n*	H1 by H2									H1 by CRASE5 (discretized)									H1 by CRASE5 (unrounded)						Degradation	
		H1		H2		Stats					H1		CRASE5		Stats					H1		CRASE5		Stats		QWK	*r*
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	M	*SD*	M	*SD*	SMD	*r*	H1CRASE5 (rounded) – H1H2	H1CRASE5 (unrounded) – H1H2
All	5,128	3.3	1.0	3.3	1.0	0.00	.56	.83	68.4	99.4	.83	3.3	1.0	3.3	1.0	0.02	.61	.85	72.0	99.6	.85	3.3	1.0	3.3	0.9	0.01	.89	0.02	0.06
Male	2,440	3.2	1.0	3.2	1.0	0.00	.59	.85	69.8	99.5	.85	3.2	1.0	3.1	1.0	0.04	.62	.86	72.7	99.7	.86	3.2	1.0	3.2	0.9	0.01	.89	0.01	0.04
Female	2,580	3.4	1.0	3.4	1.0	0.01	.54	.82	67.4	99.4	.82	3.4	1.0	3.4	0.9	0.00	.59	.84	71.4	99.5	.84	3.4	1.0	3.4	0.8	0.00	.88	0.02	0.06
Hispanic	875	3.0	1.0	3.0	1.0	0.00	.61	.85	72.2	99.5	.85	3.0	1.0	3.1	1.0	0.20*	.56	.83	68.8	99.3	.84	3.0	1.0	3.0	0.8	0.05	.88	−0.02	0.03
Not Hisp.	4,028	3.3	1.0	3.4	1.0	0.01	.54	.82	67.3	99.4	.82	3.3	1.0	3.3	1.0	0.03	.60	.84	71.7	99.6	.85	3.3	1.0	3.3	0.9	0.02	.88	0.02	0.06
Blank	1,244	2.8	1.0	2.8	1.0	0.00	.58	.83	70.1	99.3	.83	2.8	1.0	2.8	0.9	0.02	.63	.85	73.2	99.5	.85	2.8	1.0	2.8	0.9	0.03	.88	0.02	0.05
Asian	583	3.7	1.0	3.7	1.0	0.00	.51	.81	65.5	99.1	.81	3.7	1.0	3.7	1.0	0.02	.54	.82	67.4	99.3	.82	3.7	1.0	3.7	0.9	0.01	.87	0.01	0.06
Black	346	2.9	0.9	2.9	0.9	0.01	.57	.81	69.9	99.4	.81	2.9	0.9	2.8	0.9	0.05	.56	.80	70.2	99.4	.80	2.9	0.9	2.9	0.8	0.03	.87	−0.01	0.06
2+ Races	249	3.4	1.0	3.3	1.0	0.05	.55	.82	68.3	99.2	.82	3.4	1.0	3.5	1.0	0.14*	.60	.85	72.3	99.6	.86	3.4	1.0	3.3	0.9	0.06	.89	0.03	0.07
White	1,878	3.5	0.9	3.6	0.9	0.01	.50	.77	66.2	99.4	.77	3.5	0.9	3.5	0.8	0.04	.58	.80	71.6	99.6	.80	3.5	0.9	3.5	0.7	0.05	.85	0.03	0.08

H1 = human rater 1, H2 = human rater 2, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 12. ETS-Style Subgroup Analysis on Domain 3 Scores, Blind-Validation Sample

Group	*n*	H1 by H2										H1 by CRASE5 (discretized)										H1 by CRASE5 (unrounded)						Degradation	
		H1		H2		Stats						H1		CRASE5		Stats						H1		CRASE5		Stats		QWK	*r*
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	*r*	H1CRASE5 (rounded) – H1H2	H1CRASE5 (unrounded) – H1H2
All	5,128	3.4	1.0	3.4	1.0	0.00	.55	.83	68.4	99.5	.83	3.4	1.0	3.4	1.0	0.04	.60	.84	71.6	99.6	.85	3.4	1.0	3.4	0.9	0.03	.88	0.01	0.05
Male	2,440	3.3	1.0	3.3	1.0	0.00	.56	.84	68.8	99.5	.84	3.3	1.0	3.3	1.0	0.05	.61	.86	71.8	99.7	.86	3.3	1.0	3.3	0.9	0.03	.89	0.02	0.05
Female	2,580	3.5	1.0	3.5	0.9	0.01	.54	.82	68.4	99.5	.82	3.5	1.0	3.5	0.9	0.03	.59	.83	71.7	99.5	.83	3.5	1.0	3.5	0.8	0.02	.87	0.01	0.05
Hispanic	875	3.1	1.0	3.1	1.0	0.00	.58	.84	70.3	99.5	.84	3.1	1.0	3.1	1.0	0.00	.62	.85	73.0	99.5	.85	3.1	1.0	3.1	0.9	0.04	.88	0.01	0.04
Not Hisp.	4,028	3.5	1.0	3.5	1.0	0.00	.54	.82	67.9	99.5	.82	3.5	1.0	3.5	0.9	0.05	.59	.84	71.5	99.6	.84	3.5	1.0	3.5	0.8	0.04	.88	0.02	0.06
Blank	1,244	2.9	1.0	2.9	1.0	0.00	.58	.84	69.5	99.4	.84	2.9	1.0	2.9	1.0	0.04	.61	.85	72.3	99.4	.85	2.9	1.0	2.9	0.9	0.01	.88	0.01	0.04
Asian	583	3.8	1.0	3.8	0.9	0.01	.52	.81	66.9	99.5	.81	3.8	1.0	3.9	0.9	0.03	.52	.81	66.7	99.5	.81	3.8	1.0	3.8	0.8	0.01	.85	0.00	0.04
Black	346	3.0	0.9	3.0	0.9	0.02	.54	.80	67.9	99.1	.80	3.0	0.9	3.0	0.9	0.05	.57	.82	70.5	99.7	.82	3.0	0.9	3.0	0.8	0.01	.86	0.02	0.06
2+ Races	249	3.6	0.9	3.5	1.0	0.04	.56	.83	69.5	100.0	.84	3.6	0.9	3.5	1.0	0.08	.61	.85	73.1	99.6	.85	3.6	0.9	3.5	0.8	0.09	.89	0.02	0.05
White	1,878	3.7	0.8	3.7	0.8	0.01	.50	.76	67.8	99.5	.76	3.7	0.8	3.6	0.8	0.08	.57	.79	72.3	99.7	.79	3.7	0.8	3.6	0.7	0.08	.84	0.03	0.08

H1 = human rater 1, H2 = human rater 2, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 13. ETS-Style Subgroup Analysis on Domain 4 Scores, Blind-Validation Sample

Group	*n*	H1 by H2										H1 by CRASE5 (discretized)										H1 by CRASE5 (unrounded)						Degradation	
		H1		H2		Stats						H1		CRASE5		Stats						H1		CRASE5		Stats		QWK	*r*
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	*r*	H1CRASE5 (rounded) – H1H2	H1CRASE5 (unrounded) – H1H2
All	5,128	3.6	0.9	3.6	0.9	0.01	.54	.81	68.1	99.3	.81	3.6	0.9	3.6	0.9	0.04	.56	.82	69.6	99.6	.82	3.6	0.9	3.6	0.8	0.02	.86	0.01	0.05
Male	2,440	3.5	1.0	3.5	1.0	0.00	.56	.82	69.1	99.3	.82	3.5	1.0	3.5	1.0	0.07	.57	.82	69.8	99.3	.83	3.5	1.0	3.5	0.8	0.03	.86	0.00	0.04
Female	2,580	3.7	0.9	3.7	0.9	0.02	.52	.79	67.5	99.4	.79	3.7	0.9	3.7	0.9	0.02	.55	.81	69.7	99.8	.81	3.7	0.9	3.7	0.8	0.00	.85	0.02	0.06
Hispanic	875	3.3	1.0	3.3	0.9	0.01	.56	.83	69.9	99.5	.83	3.3	1.0	3.3	0.9	0.00	.56	.82	70.1	99.4	.82	3.3	1.0	3.3	0.8	0.07	.85	−0.01	0.02
Not Hisp.	4,028	3.7	0.9	3.7	0.9	0.01	.53	.79	67.8	99.3	.79	3.7	0.9	3.6	0.9	0.06	.56	.81	69.8	99.6	.81	3.7	0.9	3.7	0.8	0.04	.85	0.02	0.06
Blank	1,244	3.1	1.0	3.2	0.9	0.03	.51	.80	65.9	99.4	.80	3.1	1.0	3.1	0.9	0.05	.53	.80	67.6	99.2	.81	3.1	1.0	3.2	0.8	0.04	.84	0.00	0.04
Asian	583	4.0	0.9	4.0	0.9	0.02	.51	.78	66.6	99.1	.78	4.0	0.9	4.1	0.9	0.05	.49	.79	65.4	99.8	.79	4.0	0.9	4.0	0.8	0.02	.83	0.01	0.05
Black	346	3.2	0.9	3.2	0.9	0.02	.47	.74	64.5	98.3	.74	3.2	0.9	3.2	0.9	0.05	.56	.80	71.1	99.4	.80	3.2	0.9	3.2	0.8	0.03	.84	0.06	0.10
2+ Races	249	3.7	0.8	3.7	0.9	0.05	.53	.80	69.1	99.6	.80	3.7	0.8	3.7	0.9	0.10	.55	.80	70.3	100.0	.80	3.7	0.8	3.7	0.8	0.08	.87	0.00	0.07
White	1,878	3.9	0.8	3.9	0.8	0.00	.51	.73	69.2	99.4	.73	3.9	0.8	3.8	0.8	0.10	.54	.75	71.2	99.6	.76	3.9	0.8	3.8	0.7	0.09	.81	0.02	0.08

H1 = human rater 1, H2 = human rater 2, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 14. ETS-Style Subgroup Analysis on Domain 1 Scores (2–12 Scale), Blind-Validation Sample

Group	*n*	H1 by CRASE5 (discretized)										H1 by CRASE5 (unrounded)					
		H1		CRASE5		Stats						H1		CRASE5		Stats	
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	*r*
All	5,128	7.0	1.9	6.9	1.9	0.03	.47	.91	56.6	92.2	.91	7.0	1.9	7.0	1.8	0.02	.92
Male	2,440	6.8	2.0	6.7	2.0	0.05	.47	.91	56.2	92.1	.91	6.8	2.0	6.8	1.8	0.03	.92
Female	2,580	7.2	1.9	7.1	1.8	0.02	.47	.90	57.2	92.2	.90	7.2	1.9	7.2	1.7	0.01	.91
Hispanic	875	6.3	2.0	6.3	1.9	0.02	.50	.90	59.1	92.3	.91	6.3	2.0	6.4	1.7	0.05	.92
Not Hisp.	4,028	7.2	1.9	7.1	1.9	0.05	.46	.90	56.5	92.4	.90	7.2	1.9	7.1	1.7	0.04	.92
Blank	1,244	6.0	2.0	5.9	1.9	0.02	.47	.90	55.9	91.2	.90	6.0	2.0	6.0	1.8	0.01	.92
Asian	583	7.9	1.9	7.9	1.9	0.03	.42	.90	53.0	92.8	.90	7.9	1.9	7.9	1.7	0.01	.91
Black	346	6.1	1.8	6.1	1.8	0.02	.40	.87	52.3	92.5	.87	6.1	1.8	6.2	1.6	0.02	.89
2+ races	249	7.2	1.9	7.1	1.9	0.08	.43	.90	54.2	92.8	.90	7.2	1.9	7.1	1.7	0.07	.92
White	1,878	7.6	1.6	7.5	1.6	0.08	.47	.87	59.0	92.3	.87	7.6	1.6	7.4	1.4	0.09	.89

H1 = human rater 1, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 15. ETS-Style Subgroup Analysis on Domain 2 Scores (2–12 Scale), Blind-Validation Sample

Group	*n*	H1 by CRASE5 (discretized)										H1 by CRASE5 (unrounded)					
		H1		CRASE5		Stats						H1		CRASE5		Stats	
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	*r*
All	5,128	6.5	1.9	6.5	1.9	0.02	.50	.91	58.5	92.8	.91	6.5	1.9	6.5	1.7	0.01	.92
Male	2,440	6.4	2.0	6.3	1.9	0.03	.50	.91	59.1	92.7	.91	6.4	2.0	6.3	1.8	0.01	.93
Female	2,580	6.7	1.9	6.7	1.8	0.01	.49	.90	58.1	92.9	.90	6.7	1.9	6.7	1.7	0.00	.92
Hispanic	875	5.9	1.9	6.3	1.9	0.22*	.42	.88	52.5	89.4	.90	5.9	1.9	6.0	1.7	0.05	.92
Not Hisp.	4,028	6.7	1.9	6.6	1.9	0.03	.49	.91	58.4	92.7	.91	6.7	1.9	6.7	1.7	0.02	.92
Blank	1,244	5.5	1.9	5.5	1.8	0.00	.51	.91	59.8	93.6	.91	5.5	1.9	5.6	1.7	0.03	.92
Asian	583	7.4	1.9	7.5	1.9	0.02	.48	.90	57.6	91.6	.90	7.4	1.9	7.4	1.7	0.00	.92
Black	346	5.7	1.8	5.7	1.7	0.00	.48	.88	58.7	93.1	.88	5.7	1.8	5.8	1.6	0.04	.90
2+ races	249	6.8	1.9	7.1	1.9	0.18*	.49	.90	59.0	89.2	.91	6.8	1.9	6.7	1.7	0.04	.92
White	1,878	7.1	1.7	7.0	1.6	0.05	.46	.87	57.2	92.2	.88	7.1	1.7	7.0	1.5	0.05	.90

H1 = human rater 1, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 16. ETS-Style Subgroup Analysis on Domain 3 Scores (2–12 Scale), Blind-Validation Sample

Group	*n*	H1 by CRASE5 (discretized)										H1 by CRASE5 (unrounded)					
		H1		CRASE5		Stats						H1		CRASE5		Stats	
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	*r*
All	5,128	6.9	1.9	6.8	1.9	0.03	.48	.90	57.5	92.5	.90	6.9	1.9	6.8	1.7	0.03	.92
Male	2,440	6.7	1.9	6.6	1.9	0.04	.49	.91	58.0	92.5	.91	6.7	1.9	6.6	1.8	0.03	.92
Female	2,580	7.0	1.8	7.0	1.8	0.02	.46	.90	57.0	92.5	.90	7.0	1.8	7.0	1.6	0.02	.91
Hispanic	875	6.2	1.9	6.2	1.8	0.02	.50	.91	59.8	93.1	.91	6.2	1.9	6.3	1.7	0.04	.92
Not Hisp.	4,028	7.0	1.8	6.9	1.8	0.04	.47	.90	57.1	92.4	.90	7.0	1.8	6.9	1.7	0.04	.92
Blank	1,244	5.9	1.9	5.8	1.9	0.02	.49	.91	58.2	92.5	.91	5.9	1.9	5.9	1.7	0.01	.92
Asian	583	7.7	1.8	7.7	1.8	0.03	.43	.89	54.2	90.6	.89	7.7	1.8	7.7	1.6	0.00	.90
Black	346	6.0	1.8	6.0	1.7	0.01	.40	.88	52.9	93.1	.88	6.0	1.8	6.1	1.6	0.04	.90
2+ races	249	7.1	1.9	6.9	1.9	0.07	.46	.91	56.6	93.2	.91	7.1	1.9	6.9	1.7	0.07	.92
White	1,878	7.4	1.6	7.3	1.5	0.08	.46	.87	58.3	92.5	.87	7.4	1.6	7.3	1.4	0.09	.89

H1 = human rater 1, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

Table 17. ETS-Style Subgroup Analysis on Domain 4 Scores (2–12 Scale), Blind-Validation Sample

Group	*n*	H1 by CRASE5 (discretized)										H1 by CRASE5 (unrounded)					
		H1		CRASE5		Stats						H1		CRASE5		Stats	
		M	*SD*	M	*SD*	SMD	K	QWK	% agree.	% adj. agree.	*r*	M	*SD*	M	*SD*	SMD	*r*
All	5,128	7.3	1.8	7.2	1.8	0.04	.43	.88	54.2	90.9	.88	7.3	1.8	7.2	1.6	0.03	.90
Male	2,440	7.1	1.8	7.0	1.8	0.05	.43	.89	53.6	91.2	.89	7.1	1.8	7.0	1.6	0.04	.90
Female	2,580	7.4	1.7	7.4	1.7	0.02	.43	.87	55.1	90.7	.87	7.4	1.7	7.4	1.5	0.02	.89
Hispanic	875	6.6	1.8	6.6	1.8	0.03	.44	.88	55.4	90.3	.88	6.6	1.8	6.7	1.6	0.07	.89
Not Hisp.	4,028	7.4	1.7	7.3	1.7	0.06	.43	.88	54.5	91.2	.88	7.4	1.7	7.3	1.6	0.05	.89
Blank	1,244	6.3	1.8	6.2	1.8	0.04	.40	.87	51.7	90.2	.87	6.3	1.8	6.3	1.6	0.02	.89
Asian	583	8.1	1.7	8.2	1.8	0.05	.38	.87	50.4	91.3	.87	8.1	1.7	8.1	1.5	0.01	.89
Black	346	6.4	1.7	6.4	1.7	0.01	.41	.86	53.5	90.8	.86	6.4	1.7	6.5	1.5	0.04	.88
2+ races	249	7.5	1.8	7.4	1.7	0.06	.47	.89	58.2	91.6	.89	7.5	1.8	7.4	1.6	0.06	.91
White	1,878	7.8	1.5	7.7	1.5	0.09	.43	.84	56.8	91.6	.84	7.8	1.5	7.7	1.3	0.11*	.86

H1 = human rater 1, CRASE5 (discretized) = final score from CRASE5, CRASE5 (unrounded) = final score from CRASE5 before discretization, *n* = sample size, M = mean, *SD* = standard deviation, SMD = standardized mean difference, K = kappa, QWK = quadratic weighted kappa, % agree. = exact agreement rate, % adj. agree. = exact + adjacent agreement rate, *r* = correlation

VI. Automatic Detection of Condition Codes

In most scoring programs, the rubric includes condition codes for identifying responses that are not valid attempts at the prompt or are written in a way that makes scoring difficult or impossible. CRASE5 has new models to automatically assign condition codes to invalid responses. Since these checks happen before the response is sent to the generic scoring models, the CRASE team calls this process prescoring.

This section summarizes ACT's condition code definitions for the ACT writing test and explains how CRASE5 automatically detects responses earning a condition code.

Blanks

The response is blank.

The response is completely erased.

After whitespace characters (tabs, spaces, and return characters) are removed via string replacement, CRASE5 looks for essays with no characters. This is the same process used in CRASE+.

Voided Essays

The response is marked "void" or "voided."

CRASE5 looks for responses where every word in the response is "void," "voided," "na," or "n/a." (The checks for "na" and "n/a" were requested by ACT's Scoring Operations team in 2022.) This check is case-insensitive, and punctuation marks are ignored. This is the same process used in CRASE+.

The response is completely crossed out.

Since an examinee cannot cross out an online essay, this rule does not apply to automated scoring.

The response consists of a direct statement of refusal to participate.

We plan to use large language models (LLMs) to identify refusals to participate.

Off-Topic

Off-topic detection has been greatly enhanced in CRASE5 by using a neural-network-based encoder model to check for similarity between the prompt and response. While CRASE+ could detect only one of the four off-topic rules (a response containing a single word), CRASE5 is expected to address all four rules.

The response does not address the prompt issue or the writing task.

We use a cosine similarity score to determine the semantic similarity between the prompt and response (a response with a similarity score below 0.6 is considered an off-topic response).

The response consists solely of statements such as "I don't know" or "We haven't learned this topic."

One benefit of using a neural-network-based encoder is its generalizability. It compares the prompt and responses without a hard-coded word list and focuses on the semantic meaning of the texts. This way, we can detect various off-topic responses like the example sentences above.

The response consists of a single word.

Like CRASE+, CRASE5 can take a response, remove punctuation and numbers, and then count the number of words. If there is only one word remaining after processing, then the response is flagged as off-topic.

The response is solely a direct copy of the prompt or passage language (no sample of the student's writing is provided).

CRASE5 can not only determine whether a response does not address the prompt (the first rule for off-topic responses), it can also determine whether the response is too similar to the prompt (e.g., a direct copy). For this comparison, we use 0.9 as a cosine similarity threshold.

Illegible

The writer's intent cannot be determined because of indecipherable handwriting or other marks obscuring the writing.

Since an examinee cannot type an online essay with indecipherable handwriting, this rule does not apply to automated scoring.

If the essay is typed, random keystrokes are also assigned this code.

CRASE5 performs this check in the same way that CRASE+ did. The essay is divided into trigraphs—consecutive three-letter combinations within words. Trigraphs are compared to a list of common trigraphs (e.g., "ing," "ies," "tri"). If the percentage of common trigraphs in the essay is below a certain threshold, then the essay is flagged as illegible. Note that this approach has been used by linguists since the 1990s and is still a very common approach to detecting gibberish responses.

Additional Rules

The Scoring Operations team has asked that the CRASE illegibility checks include a check to determine whether only numbers, only punctuation, or only a combination of the two are present in a response. That is, there is no written English in the response. This was implemented in CRASE+ and is implemented in CRASE5.

Not in English

The majority of the response is in a language other than English.

Both CRASE+ and CRASE5 include a third-party Python library called *langid.py* that reads a piece of text and uses a predefined statistical model to determine the primary language of the text. If the language with the highest predicted probability is a language other than English, then the response is flagged.

Kickouts

Though the following conditions are not condition codes in ACT's scoring rules, Research and Scoring Operations determined the need for a special kickout category for two kinds of responses.

1. **Responses that contain fewer than 25 words.** Short responses can yield unpredictable automated scoring results. For example, an engine may treat a perfectly written eight-word sentence as a high-scoring essay simply because all the rules of spelling and grammar are met. CRASE5, like CRASE+, will kick out any response with fewer than 25 words, and two hand scorers will score the essay.

2. **Responses in which 20% or more of the characters are uppercase.** From previous experience, we have learned that the language identification procedure and the spell checking mechanism will return incorrect results if the response is written in all (or mostly) uppercase letters. Incorrect features will have unpredictable effects on the prediction model and may return incorrect scores. CRASE5, like CRASE+, will kick out any response where 20% or more of the characters in the essay are in uppercase. Two hand scorers will score the essay.

Disturbing Content

The response contains disturbing content.

New to CRASE5 is the ability to determine whether an essay contains text that suggests harm to the examinee, other people, or property. It is important that there is a mechanism to identify disturbing content in essays so that testing staff and school administration can follow up accordingly and as quickly as possible. Essays flagged for disturbing content will be sent to either a human hand scorer or a test administrator, who will review the essay and take proper steps to alert the proper people. Rather than using traditional rule-based methods, we fine-tuned a neural-network-based language model to classify these responses more reliably.

Summary of Prescoring Checks

For any essay submitted to the CRASE5 system, the following steps will occur:

1. CRASE5 will check whether the response is blank.

2. If the response is blank, it will receive a condition code of 1, and the CRASE5 scoring process will end. Otherwise, CRASE5 will check whether the response is non-English.

3. If the response is non-English, it will receive a condition code of 2, and the CRASE5 scoring process will end. Otherwise, CRASE5 will check whether the response is void.

4. If the response is void, it will receive a condition code of 4, and the CRASE5 scoring process will end. Otherwise, CRASE5 will check whether the response is illegible.

5. If the response is illegible, it will receive a condition code of 5, and the CRASE5 scoring process will end. Otherwise, CRASE5 will check whether the response is a kickout.

6. If the response is a kickout, it will receive a condition code of 6, and the CRASE+ scoring process will end. Otherwise, CRASE5 will check whether the response includes disturbing content.

7. If the response includes disturbing content, it will receive a condition code of 7, and the CRASE5 scoring process will end. Otherwise, CRASE5 will check whether the response is off-topic.

8. If the response is off-topic, it will receive a condition code of 3, and the CRASE5 scoring process will end.

9. If none of the prescoring models are triggered, CRASE5 will score the response on the four domains of the ACT writing rubric, and the CRASE5 scoring process will end.

10. For all cases except Steps 7 and 8, at least one hand scorer will complete the scoring process. For essays in Steps 7 and 8, at least two hand scorers will complete the scoring process.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Ramineni, C., Trapani, C. S., & Williamson, D. M. (2015). *Evaluation of e-rater for the Praxis I writing test* (Research Report RR–15-03). ETS. https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12047

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater scoring engine for the GRE issue and argument prompts* (Research Report RR–12-02). ETS. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2012.tb02284.x

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

ABOUT ACT

ACT is transforming college and career readiness pathways so that everyone can discover and fulfill their potential. Grounded in more than 65 years of research, ACT's learning resources, assessments, research, and work-ready credentials are trusted by students, job seekers, educators, schools, government agencies, and employers in the U.S. and around the world to help people achieve their education and career goals at every stage of life. Visit us at www.act.org.