CRASE+® for ACT Writing Technical Report

Scott W. Wood

I. Introduction

During 2021 and 2022, the CRASE+® research team studied the feasibility of using the CRASE+ automated scoring engine on ACT® writing tests administered online. Researchers conducted multiple proof-of-concept studies to evaluate how accurate CRASE+ scores were compared with those of human hand scorers. Additional studies looked at subgroup differences, effects on equating, and the effects of various modifications around engine training.

Based on the findings from the research studies, ACT has proposed using CRASE+ automated scoring to replace one of the two hand scorers assigned to every ACT writing test essay submitted online. Resolution reads (in which a third scorer reads the essay to resolve a difference in the scores assigned by the first two scorers) would be handled by an independent hand scorer. ACT International started using CRASE+ as one of the initial scorers this way in October 2022.

This document summarizes the primary findings from these proof-of-concept studies. The results presented here should help state assessment coordinators better understand how CRASE+ was trained and how CRASE+ scores compare with hand scores.

The next section contains a brief overview of automated scoring and the CRASE+ engine. Sections III and IV discuss the data and processes used to train the engine. Sections V and VI review subgroup analyses of the CRASE+ data. Section VII describes how condition codes are handled by CRASE+.

© 2023 by ACT, Inc. All rights reserved. | R2307

II. Background: Automated Scoring and CRASE+

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*, and 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 pre-processor 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.

CRASE+ was created in 2007 for a state's summative testing program. The system has been enhanced since then to include scoring methodologies for additional types of free-response items and to incorporate new technologies in text processing and analysis. 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.

This report assumes a basic familiarity with automated scoring concepts. For those readers that are 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://ncme.elevate.commpartners.com/products/digital-module-18-automated-scoring*
- Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press.
- 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-scoring-standards-2021-07.pdf*
- McCaffrey, D., Casabianca, 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*

© 2023 by ACT, Inc. All rights reserved. | R2307

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.

The proof-of-concept studies used ACT writing essays from three sources: the September 2020 ACT International, the October 2020 ACT International, and selected Spring 2021 State and District administrations. Only essays obtained via online administrations were included. Note that ACT National writing tests are not administered digitally at this time, so they were not included in these studies. Approximately 14,000 essays with hand scores were provided for the studies with less than 1% of the essays being excluded due to condition codes.

Table 1 shows the list of essay prompts included in the project data, along with their source administration and the number and percentage of records corresponding to each prompt. There were 27 writing prompts represented in the data. Overall, 66.8% of the records came from the State and District administration, 19.1% came from the October International administration, and 14.1% came from the September International administration.

© 2023 by ACT, Inc. All rights reserved. | R2307

Table 1. ACT Writing Prompts Used for Project

Administration	Prompt ID	Record Count	Percentage
State and District	I114_00824	2,086	14.9%
State and District	I114_00915	1,618	11.6%
State and District	I114_00789	1,447	10.3%
State and District	I114_01058	1438	10.3%
October International	I114_01170	850	6.1%
State and District	I114_00620	704	5.0%
September International	I114_01111	687	4.9%
State and District	I114_00934	662	4.7%
State and District	I114_00571	631	4.5%
October International	I114_01160	449	3.2%
State and District	I114_00939	414	3.0%
October International	I114_01167	376	2.7%
September International	I114_00921	330	2.4%
October International	I114_01171	251	1.8%
October International	I114_01169	235	1.7%
September International	I114_01120	225	1.6%
September International	I114_00972	211	1.5%
State and District	I114_00889	199	1.4%
October International	I114_00993	187	1.3%
October International	I114_00617	185	1.3%
September International	I114_01019	158	1.1%
October International	I114_01196	143	1.0%
September International	I114_01152	135	1.0%
September International	I114_01075	124	0.9%
September International	I114_01071	105	0.8%
State and District	A00343748	71	0.5%
State and District	I114_00794	69	0.5%

Table 2 gives, for (human) Rater 1, the score point distributions, means, and standard deviations for the four writing domains by administration. Examinees are most likely to receive ratings of 3 or 4 on each domain and least likely to receive ratings of 1 or 6. Also note that the September and October ACT International examinees tended to have higher scores across the four domains than the State and District examinees. This is likely due to the fact that ACT International examinees opt to take the assessment, while many State and District examinees are required to take writing for accountability purposes.

Table 2. Score Point Distributions, Means, and Standard Deviations, by Domain and Administration

Domain	Admin.	1	2	3	4	5	6	Mean	SD
1	Sept. Int.	0.7%	5.3%	21.5%	44.3%	23.2%	5.0%	3.99	0.96
	Oct. Int.	1.3%	5.0%	21.5%	44.8%	23.1%	4.3%	3.96	0.96
	S&D	3.7%	13.8%	36.3%	37.6%	8.1%	0.5%	3.34	0.96
2	Sept. Int.	0.8%	7.2%	27.0%	44.2%	18.1%	2.7%	3.80	0.94
	Oct. Int.	1.3%	7.0%	27.5%	44.0%	18.2%	2.1%	3.77	0.94
	S&D	4.0%	20.8%	39.6%	30.4%	5.1%	0.1%	3.12	0.94
3	Sept. Int.	0.7%	5.7%	23.7%	46.4%	20.3%	3.3%	3.90	0.92
	Oct. Int.	1.3%	5.3%	23.7%	47.2%	20.2%	2.3%	3.87	0.92
	S&D	3.7%	15.0%	37.7%	36.7%	6.6%	0.2%	3.28	0.93
4	Sept. Int.	0.4%	4.5%	18.0%	46.5%	25.2%	5.5%	4.08	0.93
	Oct. Int.	0.8%	4.2%	18.9%	46.5%	24.7%	4.9%	4.05	0.93
	S&D	2.2%	9.5%	37.0%	42.0%	8.8%	0.6%	3.47	0.88

Tables 3, 4, and 5 provide demographic information about the records in the project data set. Table 3 summarizes the percentage of records by reported gender. Table 4 summarizes the percentage of records by reported Hispanic status. Table 5 summarizes the percentage of records by reported race/ethnicity. When students register for the ACT, they are asked to indicate their race, marking all that apply: American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian/Other Pacific Islander, White, and Prefer not to respond or none of these apply.

The gender distribution (Table 3) is consistent across the three administrations. However, the Hispanic status distribution (Table 4) and race/ethnicity distribution (Table 5) differ between International and State and District administrations. The State and District demographics are consistent with demographics collected in the 2020 U.S. Census. The International demographics differ from the State and District demographics; note, for example, the high percentage of Asian examinees taking the ACT International administrations. Since the goal is to produce a single set of automated essay scoring models that can be applied to both State and District and International examinees, it is necessary to include a balance of both kinds of examinees to produce training and validation samples representative of the combined testing population.

Table 3. Distribution of Gender, by Administration

Administration	Male	Female	Another Gender	Prefer Not to Respond	Blank
Sept. Int.	49.4%	47.8%	0.0%	2.8%	0.0%
Oct. Int.	49.2%	48.6%	0.0%	2.1%	0.1%
S&D	47.5%	50.2%	0.5%	1.2%	0.6%

© 2023 by ACT, Inc. All rights reserved. | R2307

Table 4. Distribution of Hispanic Status, by Administration

Administration	Hispanic	Non-Hispanic	Prefer Not to Respond	Blank
Sept. Int.	4.3%	84.4%	11.3%	0.0%
Oct. Int.	4.6%	84.9%	10.2%	0.1%
S&D	19.9%	76.4%	2.2%	1.3%

Table 5. Distribution of Race/Ethnicity, by Administration

Admin.	Asian	White	Hispanic	Black	Pacific Islanders	Native American	Two or More Races	Prefer Not to Resp.	Blank
Sep. Int.	67.1%	11.3%	4.3%	1.6%	0.2%	0.2%	1.7%	13.6%	0.1%
Oct. Int.	59.2%	14.4%	4.6%	2.7%	0.1%	0.3%	1.6%	17.0%	0.1%
S&D	5.3%	35.8%	12.4%	7.8%	0.1%	0.7%	5.5%	1.5%	30.9%

Training and Validation Samples

The CRASE+ research team built *generic scoring models* for this project. A generic scoring model is 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 that prompt only. There are several advantages to using generic scoring models. One is the ability to create a small number of models and apply those models to many items. Another advantage is the ability to apply the models to new writing prompts without the need for large amounts of hand-scored field-testing data. One disadvantage of generic scoring models is the inability to leverage characteristics specific to a writing prompt.

Given the size of the prompt bank for the ACT writing test and the lack of training data for some lesser-administered prompts, the research team focused on generic scoring models during all studies.

Good statistical modeling practice states that data should be allocated to training samples and blind-validation samples. The training sample is used to determine the model of best fit. The blind-validation sample, being blind to the model-training process, is used to evaluate the model of best fit using new data.

It was important that the training and blind-validation samples had adequate representation from the ACT International and State and District administrations. We also ensured the training and blind-validation samples used different prompts. Therefore, the following rules were established for training sample and validation sample allocation:

© 2023 by ACT, Inc. All rights reserved. | R2307

1. Of the four State and District prompts with the most essays (I114_00824, I114_00915, I114_00789, and I114_01058), two were randomly selected for the training sample. The other two went to the blind-validation sample.

2. Of the seven remaining State and District prompts, four were randomly selected for the training sample. The remaining three went to the blind-validation sample.

3. The two ACT International prompts with the most essays (I114_01170 and I114_01111) were selected for the training sample.

4. Of the 14 remaining ACT International prompts, eight were randomly selected for the training sample. The remaining six went to the blind-validation sample.

Tables 6a and 6b contain information about which prompts were selected for the training and blind-validation samples, including their source administrations and the number of essays for each. The training sample contained 8,862 essays; the validation sample contained 5,128 essays.

Table 6a. Training Samples, Sorted by Record Count

Administration	Prompt ID	Record Count	Percentage
State and District	I114_00824	2,086	23.5%
State and District	I114_01058	1,438	16.2%
October International	I114_01170	850	9.6%
State and District	I114_00620	704	7.9%
September International	I114_01111	687	7.8%
State and District	I114_00571	631	7.1%
October International	I114_01160	449	5.1%
October International	I114_01167	376	4.2%
September International	I114_00921	330	3.7%
October International	I114_01171	251	2.8%
October International	I114_01169	235	2.7%
September International	I114_01120	225	2.5%
State and District	I114_00889	199	2.2%
October International	I114_00993	187	2.1%
October International	I114_01196	143	1.6%
State and District	A00343748	71	0.8%
Total	**—**	**8,862**	**100.0%**

Table 6b. Validation Samples, Sorted by Record Count

Administration	Prompt ID	Record Count	Percentage
State and District	I114_00915	1,618	31.6%
State and District	I114_00789	1,447	28.2%
State and District	I114_00934	662	12.9%
State and District	I114_00939	414	8.1%
September International	I114_00972	211	4.1%
October International	I114_00617	185	3.6%
September International	I114_01019	158	3.1%
September International	I114_01152	135	2.6%
September International	I114_01075	124	2.4%
September International	I114_01071	105	2.0%
State and District	I114_00794	69	1.3%
Total	**—**	**5,128**	**100.0%**

Tables 7, 8, and 9 contain distributions of gender, Hispanic status, and race/ethnicity by training and validation sample. While the gender and Hispanic status distributions are similar across the training and validation samples, there are some noticeable differences in the race/ethnicity distributions between the samples. This is likely due to the way that prompts from the ACT International administrations were selected. The International prompts with the largest representation appeared in the training sample. As a large percentage of ACT International examinees are Asian, this explains the discrepancy between percentage of Asian examinees in the training sample versus the validation sample.

Table 7. Distribution of Gender, by Training and Validation Sample

Sample	Male	Female	Another Gender	Prefer Not to Respond	Blank
Training	48.4%	49.2%	0.2%	1.9%	0.4%
Validation	47.6%	50.3%	0.5%	1.2%	0.4%

Table 8. Distribution of Hispanic Status, by Training and Validation Sample

Sample	Hispanic	Non-Hispanic	Prefer Not to Respond	Blank
Training	13.5%	79.6%	6.1%	0.8%
Validation	17.1%	78.5%	3.3%	1.1%

© 2023 by ACT, Inc. All rights reserved. | R2307

Table 9. Distribution of Race/Ethnicity, by Training and Validation Sample

Sample	Asian	White	Hispanic	Black	Pacific Islanders	Native American	Two or More Races	Prefer Not to Resp.	Blank
Train	31.8%	23.4%	8.8%	5.5%	0.1%	0.7%	3.8%	7.2%	18.6%
Valid	11.4%	36.6%	11.5%	6.7%	0.1%	0.2%	4.9%	4.4%	24.3%

Engine Training

Recall from Section II that features are the quantitative characteristics of a piece of text. These features get used to build a statistical model that maps text characteristics to hand scores. CRASE+ uses a default set of writing features to use for prediction. These features were developed by experts in English language arts and experts in natural language processing. In all, 39 features were available for score modeling.

CRASE+ includes many machine learning procedures to map essay features to hand scoring. In this study, gradient-boosted models were evaluated. Gradient-boosted models have historically performed well across many prompts and assessments. They use subsampling and a sequence of regression trees to build up a predictive model.

CRASE+ used *five-fold cross-validation* on the training sample to determine the best-fitting gradient-boosted regression model for scoring. In five-fold cross-validation, essays are assigned to one of five mutually exclusive groups (called *folds*). Each fold takes its turn being held out, with the remaining folds combining to form a training sample for a candidate model. The held-out fold is then scored to produce predicted scores. After each fold has been held out, the predicted scores can be used to produce accuracy and agreement metrics. This process was applied to multiple models identified by the CRASE+ researchers, and the model with the best agreement metrics was chosen as the best-fitting model.

Scoring models were trained using the final raw scores given to the examinees by hand scorers. Raw scores for the ACT writing test are determined as follows:

- If Rater 1 and Rater 2 assign the same score to an essay for a given domain, the final raw score 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 raw score 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 from 2 to 12.

Predictions from gradient-boosted regression models are decimals (for example, 8.58943820). To convert these undiscretized predictions to discretized rubric scores on the 1–6 raw scale, the researchers established cut scores based on the score point distribution of the Rater 1 scores

© 2023 by ACT, Inc. All rights reserved. | R2307

from the training set. If, for example, 8% of the training sample received a 1 from Rater 1, then the cut score was defined so that the lowest 8% of undiscretized CRASE+ scores were given a 1. If 18% of the training sample received a 2 from Rater 1, then the cut score was defined so that the next lowest 18% of undiscretized CRASE+ scores were given a 2. This procedure continued for all desired cut scores.

Engine Evaluation

Scoring models can be evaluated using both distributional metrics and agreement metrics. Distributional metrics include the score point distribution, mean, and standard deviation of the scores produced by Rater 1, Rater 2, and CRASE+. The expectation is that the CRASE+ distribution metrics are similar to those produced by Rater 1 and Rater 2.

Another key distributional metric is 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.

ACT standards require an exact agreement rate of 60% or higher. They also require that the sum of the exact and adjacent agreement rates be 95% or higher. Industry standards around automated scoring recommend that if the exact agreement rate between a hand rater and an engine does not exceed the human-human exact agreement rate, it should be within 5.125 percentage points (McGraw-Hill Education CTB, 2014).

Automated scoring professionals also report *kappa* and *quadratic weighted kappa* (QWK). These metrics, similar to correlations, are measures of rater agreement that take into account the fact that raters will sometimes agree simply by chance. Kappa incorporates penalties for disagreements. QWK incorporates greater penalties when 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, assuming that the human-human QWK also exceeds that threshold (Williamson et al., 2012). Additionally, if the human-computer QWK does not exceed the human-human QWK, it should be no more than 0.10 away (Williamson et al., 2012).

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

© 2023 by ACT, Inc. All rights reserved. | R2307

IV. Results for Engine Training and Validation

Tables 10a and 10b contain the distributional and agreement statistics for the best-fitting generic scoring model for Domain 1 (Ideas and Analysis). Statistics are based on the blind-validation sample. Statistics seen in operational practice should be comparable to those calculated for the blind-validation sample.

The Rater 1 (R1)-CRASE+ exact agreement rate exceeds 60% and exceeds the Rater 1-Rater 2 exact agreement rate. The sum of the exact and adjacent agreement rates is 99.6%, exceeding both the ACT minimum requirement and the sum of the Rater 1-Rater 2 exact and adjacent agreement rates. The Rater -CRASE+ QWK exceeds 0.70 and exceeds the Rater 1-Rater 2 QWK. Finally, the absolute standardized mean difference is between −0.15 and +0.15. Similar findings apply to the Rater -CRASE+ metrics. By all metrics and evaluation criteria, this model is appropriate for operational use.

Table 10a. Distributional Metrics, Generic Model, Domain 1 (Ideas and Analysis)

Score	Rater 1	Rater 2	CRASE+
Mean	3.5	3.5	3.4
SD	1.0	1.0	1.0
1	3.6%	3.5%	3.0%
2	11.3%	11.3%	12.6%
3	32.3%	32.9%	34.7%
4	39.0%	38.7%	37.9%
5	12.6%	12.1%	10.4%
6	1.3%	1.6%	1.3%

Note. N = 5,128

Table 10b. Agreement Metrics, Generic Model, Domain 1 (Ideas and Analysis)

Metric	Rater 1-Rater 2	R1-CRASE+	R2-CRASE+
|SMD|	0.000	0.054	0.054
SD Ratio	1.000	1.025	1.025
Exact Agree	68.0%	70.1%	69.9%
Adjacent Agree	31.3%	29.5%	29.6%
Nonadjacent Agree	0.6%	0.4%	0.5%
Kappa	.55	.58	.58
QWK	.83	.84	.84
Correlation	.83	.84	.84

Note. N = 5,128

© 2023 by ACT, Inc. All rights reserved. | R2307

Tables 11a and 11b contain the distributional and agreement statistics for the best-fitting generic scoring model for Domain 2 (Development and Support). Statistics are based on the blind-validation sample. Statistics seen in operational practice should be comparable to those calculated for the blind-validation sample.

The Rater 1-CRASE+ exact agreement rate exceeds 60% and exceeds the Rater 1-Rater 2 exact agreement rate. The sum of the exact and adjacent agreement rates is 99.6%, exceeding both the ACT minimum requirement and the sum of the Rater 1-Rater 2 exact and adjacent agreement rates. The Rater 1-CRASE+ QWK exceeds 0.70 and exceeds the Rater 1-Rater 2 QWK. Finally, the absolute standardized mean difference is between −0.15 and 0.15. Similar findings apply to the Rater 2-CRASE+ metrics. By all metrics and evaluation criteria, this model is appropriate for operational use.

Table 11a. Distributional Metrics, Generic Model, Domain 2 (Development and Support)

Score	Rater 1	Rater 2	CRASE+
Mean	3.3	3.3	3.2
SD	1.0	1.0	1.0
1	3.9%	3.8%	3.3%
2	17.6%	17.6%	18.4%
3	36.0%	35.9%	37.9%
4	33.0%	33.2%	32.5%
5	9.0%	8.7%	7.4%
6	0.5%	0.7%	0.6%

Note. N = 5,128

Table 11b. Agreement Metrics, Generic Model, Domain 2 (Development and Support)

Metric	Rater 1-Rater 2	R1-CRASE+	R2-CRASE+
|SMD|	0.003	0.030	0.033
SD Ratio	0.999	1.035	1.036
Exact Agree	68.4%	71.1%	71.7%
Adjacent Agree	31.0%	28.5%	27.8%
Nonadjacent Agree	0.6%	0.4%	0.4%
Kappa	.56	.60	.61
QWK	.83	.84	.85
Correlation	.83	.85	.85

Note. N = 5,128

© 2023 by ACT, Inc. All rights reserved. | R2307

Tables 12a and 12b contain the distributional and agreement statistics for the best-fitting generic scoring model for Domain 3 (Organization). Statistics are based on the blind-validation sample. Statistics seen in operational practice should be comparable to those calculated for the blind-validation sample.

The Rater 1-CRASE+ exact agreement rate exceeds 60% and exceeds the Rater 1-Rater 2 exact agreement rate. The sum of the exact and adjacent agreement rates is 99.7%, exceeding both the ACT minimum requirement and the sum of the Rater 1-Rater 2 exact and adjacent agreement rates. The Rater 1-CRASE+ QWK exceeds 0.70 and exceeds the Rater 1-Rater 2 QWK. Finally, the absolute standardized mean difference is between −0.15 and 0.15. Similar findings apply to the Rater 2-CRASE+ metrics. By all metrics and evaluation criteria, this model is appropriate for operational use.

Table 12a. Distributional Metrics, Generic Model, Domain 3 (Organization)

Score	Rater 1	Rater 2	CRASE+
Mean	3.4	3.4	3.4
SD	1.0	1.0	1.0
1	3.6%	3.4%	3.0%
2	12.3%	12.4%	14.2%
3	33.7%	34.0%	35.4%
4	39.0%	38.7%	38.2%
5	10.6%	10.5%	8.6%
6	0.8%	0.9%	0.7%

Note. N = 5,128

Table 12b. Agreement Metrics, Generic Model, Domain 3 (Organization)

Metric	Rater 1-Rater 2	R1-CRASE+	R2-CRASE+
|SMD|	0.002	0.060	0.062
SD Ratio	1.002	1.033	1.030
Exact Agree	68.4%	71.3%	71.2%
Adjacent Agree	31.1%	28.4%	28.5%
Nonadjacent Agree	0.5%	0.3%	0.3%
Kappa	.55	.59	.59
QWK	.83	.84	.84
Correlation	.83	.84	.84

Note. N = 5,128

© 2023 by ACT, Inc. All rights reserved. | R2307

Tables 13a and 13b contain the distributional and agreement statistics for the best-fitting generic scoring model for Domain 4 (Language Use and Conventions). Statistics are based on the blind-validation sample. Statistics seen in operational practice should be comparable to those calculated for the blind-validation sample.

The Rater 1-CRASE+ exact agreement rate exceeds 60% and exceeds the Rater 1-Rater 2 exact agreement rate. The sum of the exact and adjacent agreement rates is 99.6%, exceeding both the ACT minimum requirement and the sum of the Rater 1-Rater 2 exact and adjacent agreement rates. The R1-CRASE+ QWK exceeds 0.70 and matches the Rater 1-Rater 2 QWK. Finally, the absolute standardized mean difference is between −0.15 and 0.15. Similar findings apply to the Rater 2-CRASE+ metrics. By all metrics and evaluation criteria, this model is appropriate for operational use.

Table 13a. Distributional Metrics, Generic Model, Domain 4 (Language Use and Conventions)

Score	Rater 1	Rater 2	CRASE+
Mean	3.6	3.6	3.6
SD	0.9	0.9	0.9
1	2.0%	1.9%	1.8%
2	8.2%	8.0%	9.7%
3	31.9%	32.2%	34.0%
4	43.2%	43.3%	41.7%
5	13.3%	12.9%	11.4%
6	1.5%	1.8%	1.5%

Note. N = 5,128

Table 13b. Agreement Metrics, Generic Model, Domain 4 (Language Use and Conventions)

Metric	Rater 1-Rater 2	R1-CRASE+	R2-CRASE+
|SMD|	0.011	0.063	0.075
SD Ratio	1.003	1.006	1.003
Exact Agree	68.1%	68.6%	69.2%
Adjacent Agree	31.2%	31.0%	30.3%
Nonadjacent Agree	0.7%	0.4%	0.5%
Kappa	.54	.54	.55
QWK	.81	.81	.81
Correlation	.81	.81	.82

Note. N = 5,128

© 2023 by ACT, Inc. All rights reserved. | R2307

Overall, the four generic models performed according to ACT and automated scoring standards, as described in Section III. Because different prompts 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 14a–14k contain the Rater 1-CRASE+ standardized mean differences (SMD), exact agreement rates, and QWKs by writing domain and writing prompt. Metrics that do not meet ACT and automated scoring thresholds are indicated with an asterisk.

All but one QWK exceeded the 0.70 threshold (prompt I114_01019, Domain 4). All but four exact agreement rates exceeded the 60% threshold (prompt I114_01019, Domain 4; prompt I114_01071, Domains 1, 3, and 4). In addition, 13 of the 44 SMDs are outside the range of −0.15 and 0.15, affecting one or more domains on 5 of the 11 prompts.

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

Table 14a. R1-CRASE+ Agreement Metrics for I114_00617, by Domain (n = 185)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.24*	66.5	0.81
Development and Support	0.23*	65.9	0.79
Organization	0.25*	67.0	0.80
Language Use and Conventions	0.21*	67.0	0.80

Note. Values with asterisks represent metrics that do not meet ACT and automated scoring thresholds.

Table 14b. R1-CRASE+ Agreement Metrics for I114_00789, by Domain (n = 1,447)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.01	71.0	0.84
Development and Support	0.03	71.7	0.84
Organization	0.01	73.5	0.85
Language Use and Conventions	0.02	69.9	0.82

Table 14c. R1-CRASE+ Agreement Metrics for I114_00794, by Domain (n = 69)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.14	68.1	0.78
Development and Support	0.04	68.1	0.77
Organization	0.16*	71.0	0.78
Language Use and Conventions	0.17*	66.7	0.71

Note. Values with asterisks represents metrics that do not meet ACT and automated scoring thresholds.

© 2023 by ACT, Inc. All rights reserved. | R2307

Table 14d. R1-CRASE+ Agreement Metrics for I114_00915, by Domain (n = 1,618)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.17*	67.6	0.81
Development and Support	0.12	70.3	0.82
Organization	0.15	68.2	0.81
Language Use and Conventions	0.18*	66.8	0.78

Note. Values with asterisks represents metrics that do not meet ACT and automated scoring thresholds.

Table 14e. R1-CRASE+ Agreement Metrics for I114_00934, by Domain (n = 662)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.03	69.6	0.80
Development and Support	0.04	73.1	0.82
Organization	0.00	70.5	0.80
Language Use and Conventions	0.02	69.9	0.77

Table 14f. R1-CRASE+ Agreement Metrics for I114_00939, by Domain (n = 414)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.03	78.5	0.89
Development and Support	0.04	74.6	0.86
Organization	0.06	77.8	0.87
Language Use and Conventions	0.08	74.6	0.83

Table 14g. R1-CRASE+ Agreement Metrics for I114_00972, by Domain (n = 211)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.08	72.5	0.85
Development and Support	0.06	72.0	0.85
Organization	0.04	78.7	0.88
Language Use and Conventions	0.10	70.6	0.82

Table 14h. R1-CRASE+ Agreement Metrics for I114_01019, by Domain (n = 158)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.15	68.4	0.76
Development and Support	0.12	68.4	0.74
Organization	0.17*	69.6	0.74
Language Use and Conventions	0.23*	59.5*	0.65*

Note. Values with asterisks represents metrics that do not meet ACT and automated scoring thresholds.

Table 14i. R1-CRASE+ Agreement Metrics for I114_01071, by Domain (n = 105)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.12	59.0	0.76
Development and Support	0.20*	61.9	0.76
Organization	0.25*	57.1*	0.72
Language Use and Conventions	0.16*	55.2*	0.74

Note. Values with asterisks represents metrics that do not meet ACT and automated scoring thresholds.

Table 14j. R1-CRASE+ Agreement Metrics for I114_01075, by Domain (n = 124)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.02	77.4	0.88
Development and Support	0.04	70.2	0.84
Organization	0.04	75.8	0.86
Language Use and Conventions	0.02	74.2	0.84

Table 14k. R1-CRASE+ Agreement Metrics for I114_01152, by Domain (n = 135)

Writing Domain	|SMD|	Exact	QWK
Ideas and Analysis	0.02	72.6	0.84
Development and Support	0.01	71.1	0.83
Organization	0.02	71.9	0.82
Language Use and Conventions	0.00	65.9	0.78

© 2023 by ACT, Inc. All rights reserved. | R2307

V. Subgroup Analysis #1: Agreement Statistics

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 hand- 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” (AERA et al., 2014, 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” (AERA, 2014, 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” (AERA, 2014, 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

The following metrics are computed for reported gender (male and female), reported Hispanic status (Hispanic and non-Hispanic), and reported race/ethnicity (blank/chose not to respond, Asian, Black, two or more race/ethnicities, and White). Note that metrics for Native American students and Native Hawaiian/Other Pacific Islander students were excluded due to small sample sizes.

- 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 CRASE+ discretized scores

© 2023 by ACT, Inc. All rights reserved. | R2307

- The standardized mean difference between the Rater 1 and CRASE+ discretized scores
- The (unweighted) kappa between the Rater 1 and CRASE+ discretized scores
- The quadratic weighted kappa between the Rater 1 and CRASE+ discretized scores
- The exact agreement rate between the Rater 1 and CRASE+ discretized scores
- The sum of the Rater 1 and CRASE+ exact and adjacent agreement rates
- The Pearson correlation between the Rater 1 and CRASE+ discretized scores
- The mean and standard deviation of the Rater 1 scores (repeated again for convenience)
- The mean and standard deviation of the CRASE+ scores before they are 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 CRASE+ undiscretized scores
- The Pearson correlation between the Rater 1 scores and the CRASE+ undiscretized scores
- The Rater 1-CRASE+ (discretized) QWK minus the Rater 1-Rater 2 QWK
- The Rater 1-CRASE+ (undiscretized) Pearson correlation minus the Rater 1-Rater 2 Pearson correlation

Values in the results will be flagged and marked in boldface if

- any SMD is larger than 0.10,
- any QWK is less than 0.70,
- any exact agreement rate is less than 60%,
- any exact-plus-adjacent rate is less than 95%,
- any correlation is less than 0.70,
- the Rater 1-CRASE+ QWK minus the Rater 1-Rater 2 QWK is less than −0.10, or
- the Rater 1-CRASE+ correlation minus the Rater 1-Rater 2 Pearson correlation is less than −0.10.

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

Results

Tables 15–18 summarize the metrics described in the Methods section.

For all four domains, all gender subgroup metrics and all Hispanic/non-Hispanic subgroup metrics were within expectations. There do not appear to be any subgroup differences based on gender or Hispanic status. Across the four domains, there were four instances when a race-based metric did exceed the 0.10 threshold:

- For Domain 1, the Rater 1-CRASE+ (discretized) standardized mean difference for those identifying as multiple races was 0.12.

- For Domain 3, the Rater 1-CRASE+ (discretized) standardized mean difference for those identifying as multiple races was 0.11.
- For Domain 4, the Rater 1-CRASE+ (discretized) standardized mean difference for those identifying as multiple races was 0.12.
- For Domain 4, the Rater 1-CRASE+ (discretized) standardized mean difference for those identifying as White was 0.12.

Based on this subgroup analysis, subgroup differences are minimal and do not significantly impact scoring accuracy. There may be some mild concern regarding the standardized mean difference between the Rater 1 score and the CRASE+ score for examinees identifying as multiple races, though accuracy measures are generally unaffected.

Table 15. ETS-Style Subgroup Analysis on Domain 1 Scores, Blind-Validation Sample

Group	*n*	R1 by R2										R1 by CRASE+ (discretized)										R1 by CRASE+ (unrounded)						Degradation	
		R1		R2		Stats						R1		CRASE+		Stats						R1		CRASE+		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*	R1CRASE+ (rounded) – R1R2	R1CRASE+ (unrounded) – R1R2
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	.58	.84	70.1	99.6	.84	3.5	1.0	3.5	0.9	0.03	.88	0.01	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	0.07	.58	.85	70.0	99.6	.85	3.4	1.0	3.4	0.9	0.04	.89	0.00	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	3.5	0.9	0.03	.58	.83	70.3	99.6	.83	3.6	1.0	3.6	0.8	0.01	.87	0.01	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	3.2	1.0	0.00	.60	.84	71.4	99.4	.85	3.2	1.0	3.2	0.9	0.04	.87	−0.01	0.02
Non-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	0.07	.58	.84	69.9	99.7	.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	0.05	.58	.84	69.6	99.5	.84	3.0	1.0	3.0	0.9	0.01	.87	0.00	0.03
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	0.03	.54	.83	67.8	99.5	.83	3.9	1.0	3.9	0.8	0.01	.86	0.01	0.04
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	0.05	.57	.82	70.8	99.4	.83	3.1	1.0	3.1	0.8	0.01	.85	0.04	0.07
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	0.12*	.58	.84	70.7	99.6	.85	3.6	0.9	3.5	0.9	0.09	.89	0.02	0.07
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	0.10	.56	.79	71.0	99.8	.79	3.8	0.9	3.7	0.7	0.09	.84	0.02	0.07

Note. Values with asterisks represent metrics that do not meet ACT and automated scoring thresholds.
R1 = Rater 1, R2 = Rater 2, CRASE+ (discretized) = final score from CRASE+, CRASE+ (unrounded) = final score from CRASE+ 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 2 Scores, Blind-Validation Sample

Group	*n*	R1 by R2										R1 by CRASE+ (discretized)										R1 by CRASE+ (unrounded)						Degradation	
		R1		R2		Stats						R1		CRASE+		Stats						R1		CRASE+		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*	R1CRASE+ (rounded) – R1R2	R1CRASE (unrounded) – R1R2
All	5,128	3.3	1.0	3.3	1.0	0.00	.56	.83	68.4	99.4	.83	3.3	1.0	3.2	1.0	0.03	.60	.84	71.1	99.6	.85	3.3	1.0	3.3	0.9	0.01	.88	0.01	0.05
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	0.05	.61	.86	71.8	99.8	.86	3.2	1.0	3.2	0.9	0.02	.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	0.01	.58	.83	70.5	99.5	.83	3.4	1.0	3.4	0.8	0.01	.87	0.01	0.05
Hispanic	875	3.0	1.0	3.0	1.0	0.00	.61	.85	72.2	99.5	.85	3.0	1.0	3.0	0.9	0.01	.61	.84	72.0	99.4	.84	3.0	1.0	3.0	0.8	0.04	.87	−0.01	0.02
Non-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	0.04	.59	.84	70.9	99.7	.84	3.3	1.0	3.3	0.8	0.03	.88	0.02	0.06
Blank	1,244	2.8	1.0	2.8	1.0	0.0	.58	.83	70.1	99.3	.83	2.8	1.0	2.7	0.9	0.04	.60	.84	71.2	99.6	.84	2.8	1.0	2.8	0.8	0.03	.87	0.01	0.04
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	.53	.82	66.4	99.5	.82	3.7	1.0	3.7	0.8	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	0.03	.56	.81	70.2	99.7	.81	2.9	0.9	2.9	0.8	0.03	.86	0.00	0.05
2+ Races	249	3.4	1.0	3.3	1.0	0.05	.55	.82	68.3	99.2	.82	3.4	1.0	3.3	1.0	0.06	.64	.86	75.1	99.6	.86	3.4	1.0	3.3	0.8	0.08	.89	0.04	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	0.06	.57	.80	71.5	99.7	0.8	3.5	0.9	3.5	0.7	0.06	.85	0.03	0.08

Note. R1 = Rater 1, R2 = Rater 2, CRASE+ (discretized) = final score from CRASE+, CRASE+ (unrounded) = final score from CRASE+ 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

© 2023 by ACT, Inc. All rights reserved. | R2307

Table 17. ETS-Style Subgroup Analysis on Domain 3 Scores, Blind-Validation Sample

Group	*n*	R1 by R2										R1 by CRASE+ (discretized)										R1 by CRASE+ (unrounded)						Degradation	
		R1		R2		Stats						R1		CRASE+		Stats						R1		CRASE+		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*	R1CRASE+ (rounded) – R1R2	R1CRASE+ (unrounded) – R1R2
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	0.06	.59	.84	71.3	99.7	.84	3.4	1.0	3.4	0.8	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	0.08	.60	.85	71.3	99.9	.86	3.3	1.0	3.3	0.9	0.04	.89	0.01	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	71.5	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	0.9	0.02	.60	.84	72.0	99.4	.84	3.1	1.0	3.1	0.8	0.03	.87	0.00	0.03
Non-Hisp.	4,028	3.5	1.0	3.5	1.0	0.00	.54	.82	67.9	99.5	.82	3.5	1.0	3.4	0.9	0.07	.59	.84	71.2	99.8	.84	3.5	1.0	3.5	0.8	0.05	.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	0.9	0.06	.62	.85	72.5	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	3.8	0.9	0.00	.51	.81	66.2	99.8	.81	3.8	1.0	3.8	0.8	0.00	.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	0.06	.55	.81	69.9	100.0	.81	3.0	0.9	3.0	0.8	0.00	.85	0.01	0.05
2+ Races	249	3.6	0.9	3.5	1.0	0.04	.56	.83	69.5	100.0	.84	3.6	0.9	3.4	1.0	0.11*	.62	.85	73.5	99.6	.86	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	0.10	.56	.79	72.0	99.8	.79	3.7	0.8	3.6	0.7	0.10	.84	0.03	0.08

Note. Values with asterisks represent metrics that do not meet ACT and automated scoring thresholds.
R1 = Rater 1, R2 = Rater 2, CRASE+ (discretized) = final score from CRASE+, CRASE+ (unrounded) = final score from CRASE+ 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 18. ETS-Style Subgroup Analysis on Domain 4 Scores, Blind-Validation Sample

Group	*n*	R1 by R2										R1 by CRASE+ (discretized)										R1 by CRASE+ (unrounded)						Degradation	
		R1		R2		Stats						R1		CRASE+		Stats						R1		CRASE+		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*	R1CRASE+ (rounded) – R1R2	R1CRASE+ (unrounded) – R1R2
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	0.06	.54	.81	68.6	99.6	.81	3.6	0.9	3.6	0.8	0.03	.85	0.00	0.04
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	0.09	.55	.82	68.8	99.5	.83	3.5	1.0	3.5	0.8	0.05	.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	0.03	.53	.80	68.4	99.7	.80	3.7	0.9	3.7	0.8	0.00	.85	0.01	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	0.02	.53	.80	67.8	99.2	.81	3.3	1.0	3.3	0.8	0.06	.84	−0.03	0.01
Non-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	0.08	.54	.81	69.0	99.7	.81	3.7	0.9	3.7	0.8	0.05	.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	0.07	.51	.80	66.1	99.4	.80	3.1	1.0	3.2	0.8	0.03	.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	0.04	.48	.79	64.3	99.8	.79	4.0	0.9	4.0	0.7	0.01	.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	0.06	.54	.79	69.4	99.7	.79	3.2	0.9	3.2	0.7	0.02	.84	0.05	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	3.6	0.9	0.12*	.54	.80	69.9	100.0	.81	3.7	0.8	3.7	0.8	0.10	.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	0.12*	.54	.75	71.2	99.6	.76	3.9	0.8	3.8	0.6	0.10	.81	0.02	0.08

Note. Values with asterisks represent metrics that do not meet ACT and automated scoring thresholds.
R1 = Rater 1, R2 = Rater 2, CRASE+ (discretized) = final score from CRASE+, CRASE+ (unrounded) = final score from CRASE+ 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. Subgroup Analysis #2: Essay Features

The previous section focused on domain scores, whether from hand scoring or from the CRASE+ automated scoring models. Because automated scores are a weighted sum of essay features, it is appropriate to perform subgroup analyses on the features themselves. Researchers have proposed using modified differential item functioning techniques to study subgroup differences. Called *differential feature functioning* (DFF), these techniques identify differences in average feature value, conditional on the domain score, for each subgroup.

Two approaches to differential feature functioning are available: graphical and numeric. In the graphical approach, referred to as a *conditional DFF plot*, the domain score is plotted on the horizontal axis, and the average feature for a subgroup is plotted on the vertical axis. To place features on the same scale for comparison purposes, Zhang et al. (2017) recommend transforming a feature using the function

$$Y = \frac{Z - \min(Z)}{\max(Z) - \min(Z)},$$

where *Z* is the feature value and *Y* is the transformed value. The average feature is then calculated using these transformed values.

To minimize the impact of errors that may be present in CRASE+ domain scores, domain scores from Rater 1 will be used along the horizontal axis.

If the subgroup lines overlap along the domain scores, then there is little evidence of subgroup differences. If the subgroup lines diverge along the domain of the scores, then there is evidence of subgroup differences.

The second approach to DFF is numeric DFF. In this approach, a statistic related to the differential item functioning standardization statistic is computed using the function

$$STD\text{-}EISDIF = \frac{\sum_{m=1}^{M} N_{fm}\left(E_f(Y|X = m) - E_r(Y|X=m)\right)}{\sum_{m=1}^{M} N_{fm}},$$

where *M* is the number of score points used for the domain score, N_{fm} is the number of examinees in the focal subgroup with the m^{th} domain score, and $E(Y \mid X = m)$ is the average transformed feature value for examinees with the m^{th} domain score and the appropriate subgroup (*f* = focal group, *r* = reference group). The larger the value of *STD-EISDIF*, the more likely subgroup differences exist. Zhang et al. (2017) obtained *STD-EISDIF* statistics less than 0.05 in an analysis based on GRE Analytical Writing essays.

Methods

Prior to DFF analysis, the CRASE+ research team determined the seven features with the highest relative importance across the four domains. *Relative importance* is a measure of how valuable a feature is to a gradient-boosted regression model. The features most predictive in a model have the highest relative importance.

For intellectual property reasons, the seven features cannot be listed or described in this report. They will be labeled Feature 1 through Feature 7, with Feature 1 having the highest relative importance, Feature 2 having the next highest importance, and so on.

Five focal/reference group pairings were considered in this analysis:

- Gender, with male as the reference group and female as the focal group
- Hispanic status, with non-Hispanic as the reference group and Hispanic as the focal group
- Race, with White as the reference group and blank/prefer not to respond as the focal group
- Race, with White as the reference group and non-White (i.e., Native American, Asian, Black, Native Hawaiian/Other Pacific Islander, and Hispanic students) as the focal group
- Race, with blank/prefer not to respond as the reference group and non-White (as defined in the previous bullet) as the focal group

The non-White subgroup was used because the individual race/ethnicity categories had small enough sample sizes to affect the final results.

For all graphical DFF plots, only Domain 1 scores will be used along the horizontal axes. The plots looked very similar for all four domains.

For numeric DFF tables, any values greater than 0.05 or less than 0.05 will be flagged.

Results

Table 19 contains the numeric DFF results based on gender. Figures 1 and 2 illustrate the graphical DFF plots for each of the seven features studied based on gender subgroups. Since all DFF values are less than 0.05 and the lines in the graphics generally overlap, it is concluded that subgroup differences are minimal.

Table 19. Differential Feature Functioning Results Using Gender Subgroups (Reference Group: Male; Focal Group: Female)

Feature	Domain 1	Domain 2	Domain 3	Domain 4
1	0.022	0.021	0.021	0.024
2	0.018	0.017	0.018	0.020
3	−0.004	−0.004	−0.004	−0.004
4	0.004	0.004	0.004	0.004
5	0.011	0.010	0.010	0.011
6	0.004	0.004	0.004	0.004
7	−0.006	−0.005	−0.006	−0.008

Figure 1. Differential Feature Functioning Graphs for Features 1–4, Using Gender Subgroups

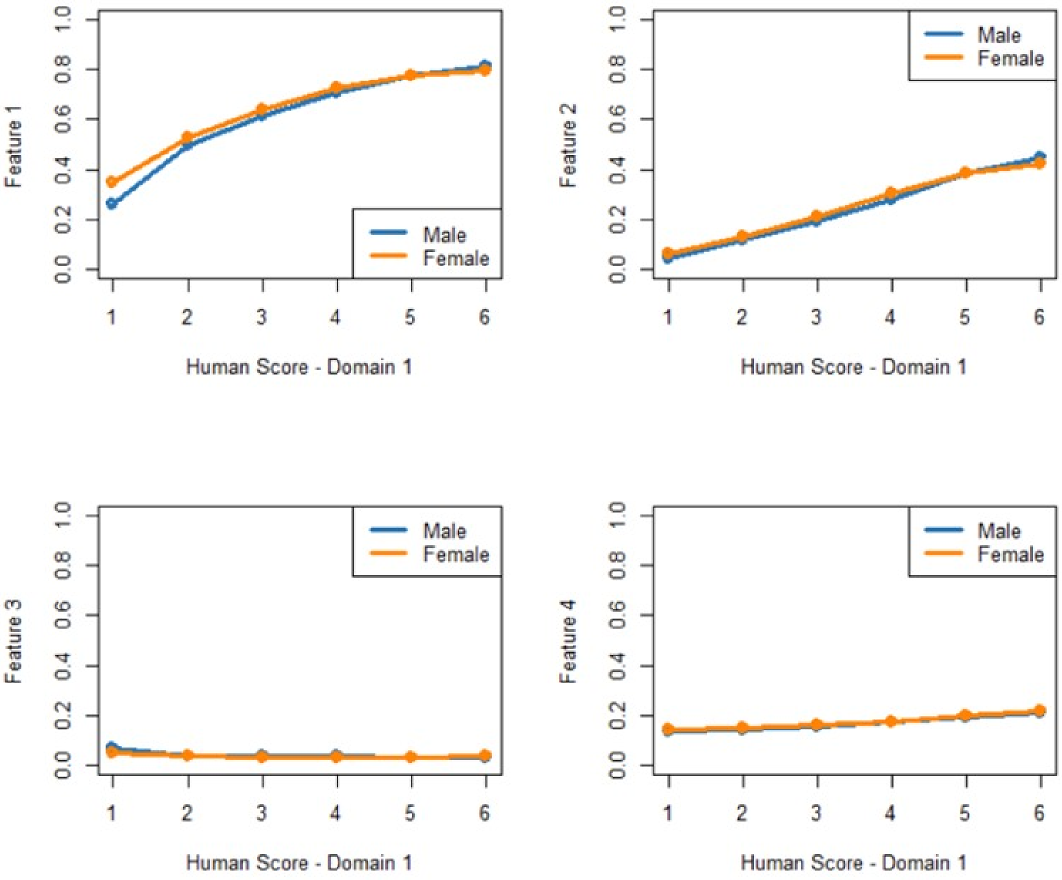


Figure 2. Differential Feature Functioning Graphs for Features 5–7, Using Gender Subgroups

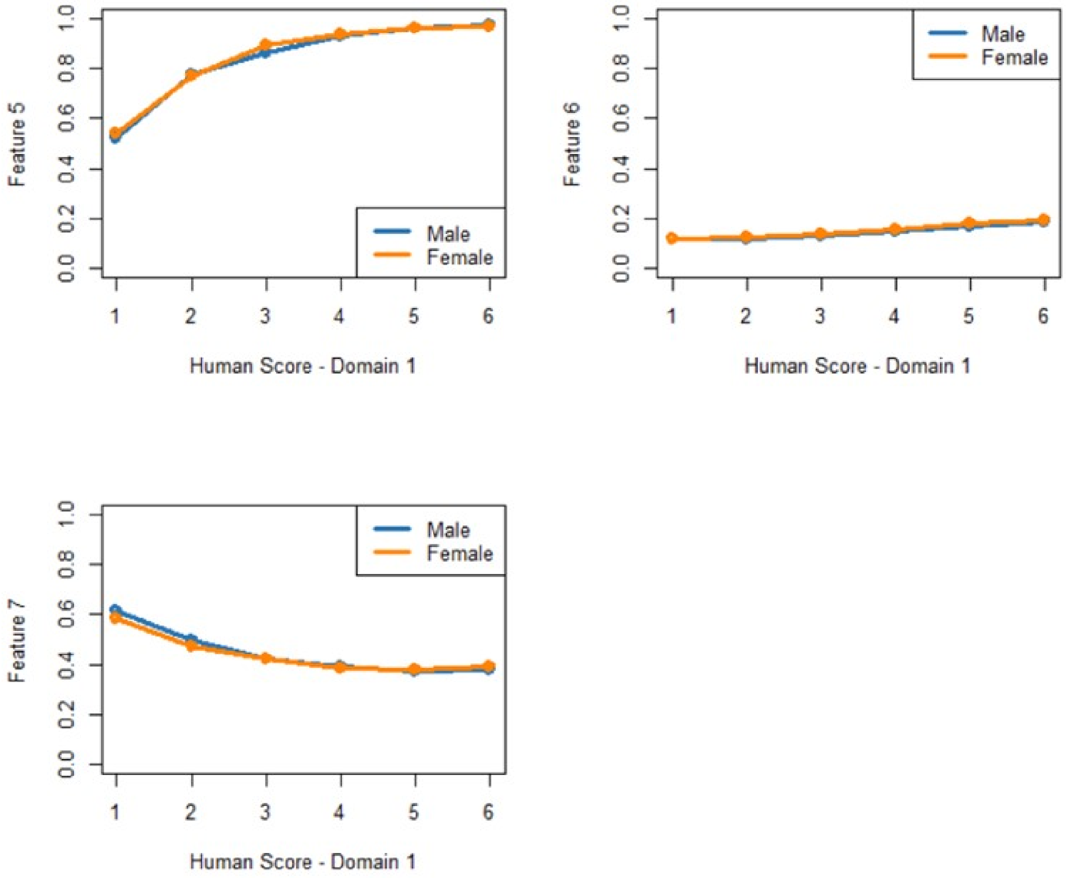


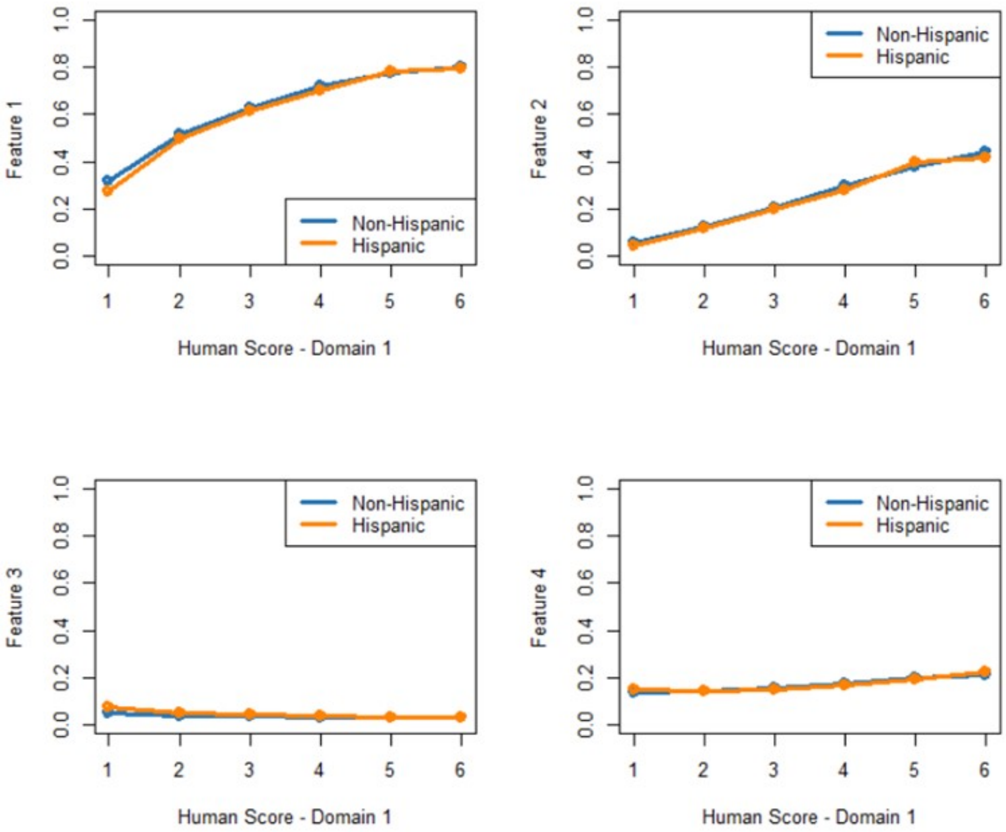
Table 20 contains the DFF calculations based on Hispanic status. Figures 3 and 4 illustrate the expected conditional means for each of the seven features for Hispanic and non-Hispanic subgroups. All DFF statistics are less than 0.05 in magnitude. All Hispanic and non-Hispanic lines in the graphics appear to overlap. Therefore, subgroup differences are minimal.

© 2023 by ACT, Inc. All rights reserved. | R2307

Table 20. Differential Feature Functioning Results Using Hispanic Subgroups (Reference Group: Non-Hispanic; Focal Group: Hispanic)

Feature	Domain 1	Domain 2	Domain 3	Domain 4
1	−0.018	−0.020	−0.019	−0.014
2	−0.010	−0.012	−0.010	−0.009
3	0.007	0.007	0.007	0.006
4	−0.004	−0.005	−0.005	−0.003
5	−0.017	−0.019	−0.017	−0.009
6	−0.006	−0.007	−0.006	−0.005
7	−0.011	−0.010	−0.010	−0.011

Figure 3. Differential Feature Functioning Graphs for Features 1–4, Using Hispanic Subgroups



© 2023 by ACT, Inc. All rights reserved. | R2307

Figure 4. Differential Feature Functioning Graphs for Features 5–7, Using Hispanic Subgroups

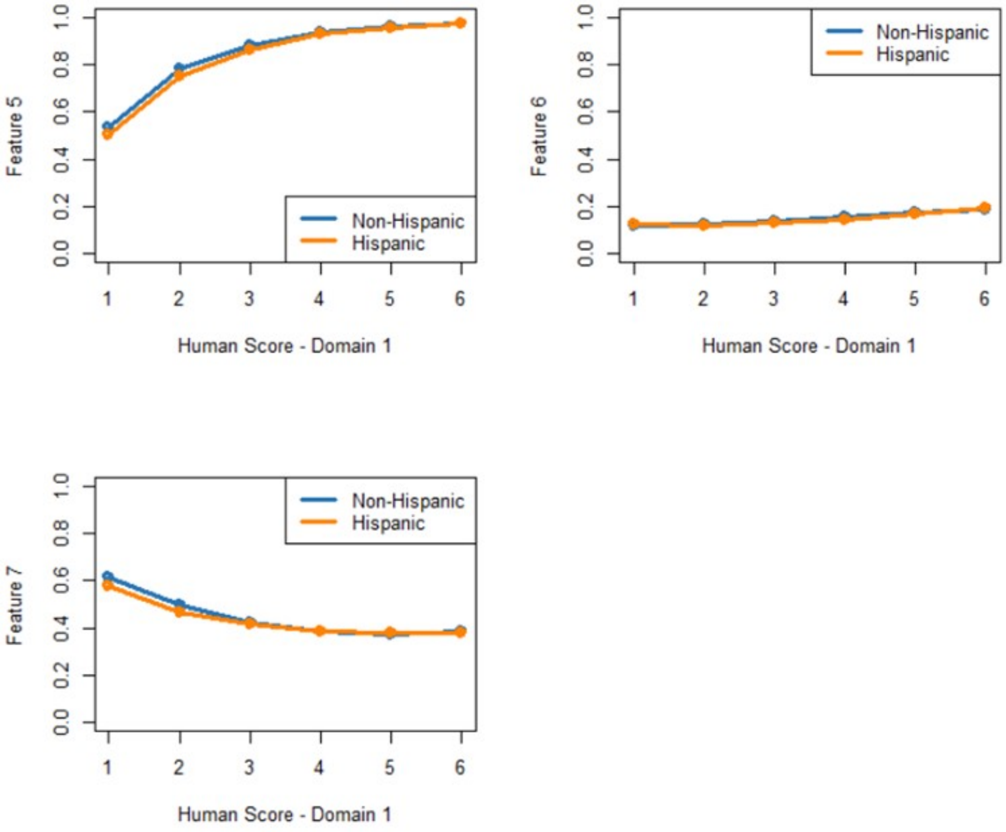
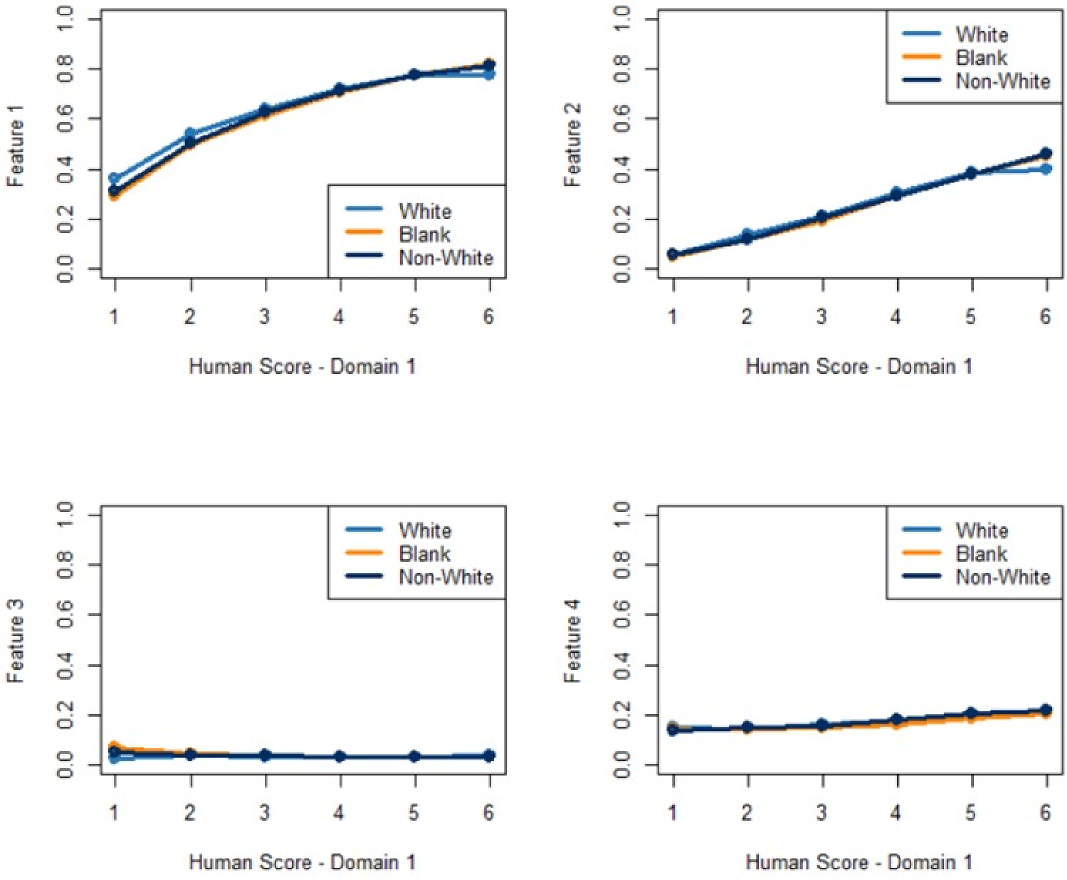


Table 21 contains the DFF calculations based on race/ethnicity, with White examinees as the reference group and blank/prefer not to respond as the focal group. Figures 5 and 6 illustrate the expected conditional means for each of the seven features for White, blank/prefer not to respond, and non-White subgroups. All DFF statistics are less than 0.05 in magnitude. All lines in the graphics appear to overlap. Therefore, subgroup differences are minimal.

Table 21. Differential Feature Functioning Results Using Race/Ethnicity Subgroups (Reference Group: White; Focal Group: Blank/Prefer Not to Respond)

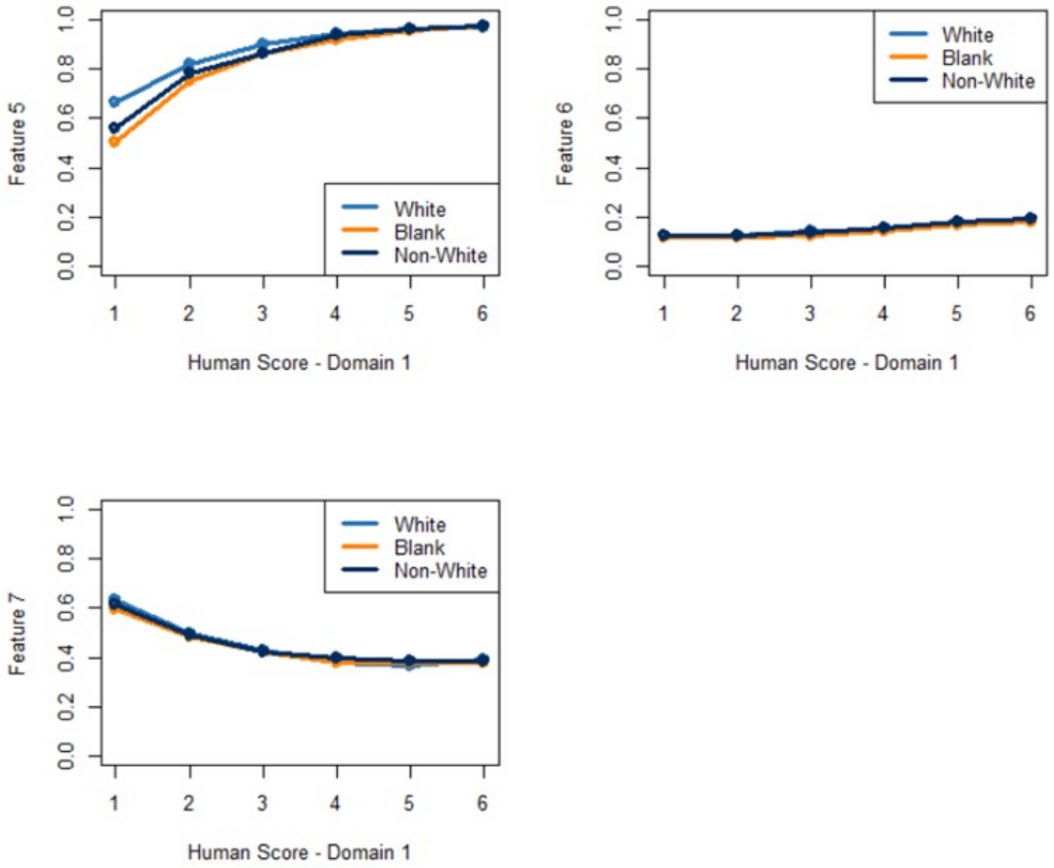
Feature	Domain 1	Domain 2	Domain 3	Domain 4
1	−0.027	−0.030	−0.028	−0.026
2	−0.014	−0.016	−0.014	−0.014
3	0.014	0.015	0.014	0.013
4	−0.027	−0.029	−0.028	−0.022
5	−0.046	−0.051	−0.048	−0.033
6	−0.036	−0.038	−0.037	−0.029
7	−0.006	−0.004	−0.005	−0.002

Figure 5. Differential Feature Functioning Graphs for Features 1–4, Using Race/Ethnicity Subgroups



© 2023 by ACT, Inc. All rights reserved. | R2307

Figure 6. Differential Feature Functioning Graphs for Features 5–7, Using Race/Ethnicity Subgroups



VII. 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. CRASE+ has 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 *pre-scoring*.

This section summarizes ACT's condition code definitions for the ACT writing test, how CRASE+ automatically detects responses earning a condition code, and how often such responses appear in the data used for model training and validation.

Note that some condition codes are not identifiable by CRASE+ at this time. These will be identified in this section. It is assumed that the hand scorer (Rater 2) will review essays for such condition codes.

Blank

The response is blank.

After whitespace characters (tabs, spaces, return characters) are removed via string replacement, CRASE+ looks for essays with no characters. This identifies all responses that are of character length zero, along with responses where the examinee typed only characters that would be invisible to a hand scorer.

To test this, we took the 3,081 project records receiving a writing condition code of 1 and ran them through the CRASE+ blank check. The blank check correctly identified the 3,075 records with blank essays. (The six remaining records were found to contain non-whitespace characters and should not have been assigned a writing condition code of 1 by hand scorers.)

The response is completely erased.

The above check should also satisfy this definition of a blank response.

Voided Essay

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

CRASE+ identifies a response as void if 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.) This check is case-insensitive, meaning that "VOID," "void," and "VoId" would each be assigned a condition code. Punctuation marks are ignored.

There were no responses in the project data set that were marked "void," "voided," "na," "n/a," or any uppercase/lowercase combination of such markings.

© 2023 by ACT, Inc. All rights reserved. | R2307

The response is completely crossed out.

Since the examinee is unable to cross out an online essay, this characteristic is not applicable to pre-scoring.

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

Unfortunately, CRASE+ is not configured to accurately identify such a voided essay. Therefore, Scoring Operations has requested that hand scorers be on the lookout for such essays.

Off-Topic

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

CRASE+ is not configured to accurately identify such off-topic responses. (The CRASE+ research team is investigating how to include this functionality in a future CRASE+ update.) Hand scorers will be on the lookout for such essays.

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

Hand scorers will be responsible for identifying such responses during operational scoring.

The response consists of a single word.

The CRASE+ system takes a response, removes punctuation and numbers, and then counts the number of words in the processed response.

In the project data set, there were 10 essays that would have received an off-topic condition code from CRASE+ for being a single word. All 10 essays had also received the off-topic condition code from hand scorers for being a single word.

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

Although CRASE+ has functionality to check for this, the models require prompt-specific modeling instead of the generic score modeling developed in these projects. Hand scorers will be responsible for identifying such responses during operational scoring.

Illegible

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

Since typewritten responses cannot be deemed indecipherable handwriting, this characteristic is not applicable to CRASE+ pre-scoring.

© 2023 by ACT, Inc. All rights reserved. | R2307

If typed, random keystrokes are also assigned this code.

CRASE+ establishes whether a response consists mainly of random keystrokes by identifying trigraphs—consecutive three-letter combinations within words. Some trigraphs, like “ing,” “ies,” and “tri,” appear frequently in English-language words. Others, like “zzz” and “qwz,” do not appear in English at all. This process, used by linguists since the 1990s, determines the trigraphs that make up an essay and compares them to a list of common English trigraphs. The higher the percentage of essay trigraphs that are common English trigraphs, the more likely the essay is written in English and the less likely the essay is nonsensical.

Running the project data through this CRASE+ module reveals that two essays would have been flagged as illegible due to random keystrokes: “yessss hhaaaaaa” and “zamalekkkkkkkkkkkkkkkkkkkkkkk.” Hand scorers did not consider either response to be random keystrokes. By adjusting the threshold used in the CRASE+ model, we can make the scoring model more conservative or liberal in its assigning of this condition code.

Additional rules.

The scoring operations team asked that CRASE+’s 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.

Not in English

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

CRASE+ contains a third-party library called *langid.py* that reads a piece of text and uses a predefined statistical model to determine the primary language of the text. The statistical model calculates the probability of a piece of text being in the languages represented in the library and returns the language with the highest probability.

To test this, we took the entire project data set and ran the data through the langid.py library. CRASE+ identified four records as not in English. CRASE+ identified the two records receiving the Not in English condition code from hand scorers—both written entirely in Spanish. A third record identified by CRASE+ was written in all capital letters. Responses written entirely in capital letters are identified by langid.py as being in other languages even if the text itself is in English. (See the next section for a resolution to this issue.) The fourth record identified by CRASE+ contained some nonsensical language, which langid.py believed was a language other than English.

Kickouts

Though this is not a condition code in ACT’s scoring rules, Research and Scoring Operations determined the need for a special kickout category for two kinds of responses with which CRASE+ (and most automated scoring engines) might struggle.

© 2023 by ACT, Inc. All rights reserved. | R2307

1. **Responses that contain fewer than 25 words.** Short responses can yield unpredictable results in an automated scoring engine. 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. CRASE+ will assign any response with fewer than 25 words (aside from zero-word and one-word responses) to the kickout category. It is expected that two hand scorers will score these short responses.

2. **Responses in which 20% or more of the characters are uppercase.** As reported in the Not-in-English section above, the language-identification program can return incorrect results if the response is written in all (or mostly all) uppercase letters. Similarly, some CRASE+ essay scoring functions will also return incorrect results for a response in all capital letters, as such responses may be viewed as a sequence of acronyms and abbreviations. This is more common in the scoring of elementary and middle school essays, but in the rare case this happens on the ACT writing test, CRASE+ will check for this and kick out the response if there are too many uppercase letters in the essay. It is expected that two hand scorers will score these unique responses.

Summary of Pre-Scoring Checks

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

1. CRASE+ will check whether the response is blank.
2. If the response is blank, it will receive a condition code of 1, and the CRASE+ scoring process will end. Otherwise, CRASE+ will check whether the response is void.
3. If the response is void, it will receive a condition code of 4, and the CRASE+ scoring process will end. Otherwise, CRASE+ will check whether the response is off topic (that is, a single word).
4. If the response is off topic, it will receive a condition code of 3, and the CRASE+ scoring process will end. Otherwise, CRASE+ will check whether the response is illegible.
5. If the response is illegible, it will receive a condition code of 5, and the CRASE+ scoring process will end. Otherwise, CRASE+ will check whether the response is non-English.
6. If the response is non-English, it will receive a condition code of 2, and the CRASE+ scoring process will end. Otherwise, CRASE+ will check whether the response is a kickout.
7. If the response is a kickout, it will receive a condition code of 6, and the CRASE+ scoring process will end.
8. If none of the pre-scoring models are triggered, CRASE+ will score the response on the four domains of the ACT writing rubric, and the CRASE+ scoring process will end.
9. For all cases except Step 7, at least one hand scorer will complete the scoring process. For essays in Step 7, at least two hand scorers will complete the scoring process.

© 2023 by ACT, Inc. All rights reserved. | R2307

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*

McGraw-Hill Education CTB. (2014). *Smarter Balanced Assessment Consortium: Field test: Automated scoring research studies in accordance with Smarter Balanced RFP 17.* Smarter Balanced Assessment Consortium.

Ramineni, C., Trapani, C. S., & Williamson, D. M. (2015). *Evaluation of* e-rater® *for the* Praxis I® *writing test*. 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*. 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.

Zhang, M., Dorans, N., Li, C., & Rupp, A. (2017). Differential feature functioning in automated essay scoring. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 185–208). Information Age Publishing.

© 2023 by ACT, Inc. All rights reserved. | R2307

ABOUT ACT

ACT is a mission-driven, nonprofit organization dedicated to helping people achieve education and workplace success. Grounded in more than 60 years of research, ACT is a trusted leader in college and career readiness solutions. Each year, ACT serves millions of students, job seekers, schools, government agencies, and employers in the U.S. and around the world with learning resources, assessments, research, and credentials designed to help them succeed from elementary school through career.

For more information, visit act.org

© 2023 by ACT, Inc. All rights reserved. | R2307