GMAC®

Graduate
Management
Admission
Council®
Creating Access to Graduate Business Education®

GMAT® Analytic Rubric Study Report

GMAC® Research Reports • RR-06-04 • March 31, 2006

The following report is the result of a special study to develop and pilot test an analytic scoring rubric for the GMAT® Analytical Writing Assessment (AWA) Analysis of an Argument prompt. GMAT® AWA has been designed as a direct measure of an examinee's ability to think critically and communicate ideas.

The study included the development of an analytic rubric that recognized the same characteristics of writing that are currently featured in the AWA Analysis of an Argument holistic writing rubric. Four distinct analytic domains were defined. The analytic rubric was applied to a small sample of responses using two different methodologies (Stage 1). Based on the results of Stage 1, the newly developed rubric was applied to a set of responses by trained human readers and by Vantage Learning Inc.'s Automated Essay Scoring (AES) system. Scores from the analytic rubric were compared to the previously assigned holistic scores. Results from the analytic rubric and the holistic rubric were compared in terms of scoring accuracy statistics and distributional characteristics (Stage 2).

Trained human readers were able to consistently apply the analytic scoring rubric across all domains. The AES provided less consistent results in the application of the analytic scoring rubric, particularly in domains that emphasized logical analysis and critical reasoning. Results from the analytic scoring were highly correlated with the holistic rubric across domains.

Given GMAC®'s interest in providing additional information to users of the AWA, the analytic scoring approach was able to provide four reliable scores that represented four domains of writing. The analytic scores provided more detailed information about the examinee's ability to write. This additional information may be useful in making decisions about applicants.

Background

The GMAT® Analytical Writing Assessment (AWA) was designed as a direct measure of an examinee's ability to think critically and communicate ideas. The AWA currently consists of two 30-minute writing tasks: Analysis of an Issue (ISS) and Analysis of an Argument (ARG). The correlation between these two tasks was high (.80) and may be as high as .96 when corrected for unreliability. Given this strong relationship between tasks, GMAC® wanted to explore alternative models that may provide more valuable and efficient information to its constituency. GMAC® commissioned ACT Inc. to consider the implications of changing the Analysis of an Argument scoring rubric from the current holistic scoring rubric to an analytic scoring rubric. The purpose was to determine if the analytic rubric would provide additional information to end users of the GMAT® AWA test section. The purpose of this document is to report the findings of the Analytic Rubric study.

Rubric Design

The first task was to convert the existing holistic Analysis of an Argument Scoring Rubric into an Analytic Rubric. Using the characteristics of writing assessed by the existing six-point holistic rubric as a starting point, ACT designed four separate analytic rubric domains, each of which was to be scored on a scale ranging from a low score of one to a high score of six. (Detailed score point descriptions are provided in Appendix A.) ACT's original proposal suggested the possibilities of identifying five possible domains from the current holistic rubric. However, upon further articulation and review of these domains, only four distinct domains could be identified.

Draft versions of the domains were reviewed by external writing content consultants. Based on the results of this review, ACT revised the domains. The domains are:

- **Domain A:** Identifies and analyzes significant flaws in the argument
- **Domain B:** Supports the critique using relevant supporting reasons and/or examples
- **Domain C:** Develops a clearly organized and coherent response
- **Domain D:** Demonstrates control of language, including diction, syntax, and conventions of standard written English

Once the analytic domains and the scoring rubric were finalized, rangefinding was conducted using the analytic domains. Training materials were assembled for each of the four analytic scales to be used with three operational Argument (ARG) prompts. These three prompts were selected to represent a variety of topics and arguments.

Methodology – Stage 1

Determining the Analytic Scoring Method

In the first stage of the study, ACT investigated whether there was a "halo effect" in the assignment of analytic scores by a single reader after one reading. That is, would readers tend to apply uniform scores across the four analytic scores to an essay (e.g., 4/4/4/4 on an "adequate" essay), or would they show a desirable willingness to assign varying domain scores based on varying levels of ability in each domain?

To determine whether a "halo effect" was occurring, 150 responses were randomly chosen from a pool of responses with existing ETS-assigned holistic scores for three operational ARG prompts (50 per prompt), according to a fixed bell curve distribution of scores.

Two different analytic scoring methods were used in this stage of the study. Prior to scoring essays using each method, experienced raters were trained in the use of the four analytic rubrics.

- **Method 1:** Three experienced readers were instructed to assign scores in all four domains to each essay based on one reading. Two independent ratings of four analytic scores were applied to each essay.
- **Method 2:** The same three readers scored the responses in only one domain for each reading: all 50 responses on each prompt were scored on the basis of a single domain before proceeding to the next domain. Thus, each response was read four times to get one full set of analytic scores. Two ratings of four analytic scores were applied to each essay, for a total of eight independent readings per response, randomly distributed among the three readers.

The results of the two different approaches were compared to determine if there was a confounding "halo effect" in the assignment of analytic scores by a single reader on one reading and to determine if one method of scoring was appreciably more consistent than the other. ACT also investigated the question, "Does one analytic scoring method yield more accurate scoring results than the other, as measured by inter-rater agreement rates and Pearson-r correlation?"

Data Analysis and Findings – Stage 1

Comparison of Method 1 and Method 2 for Assigning Analytic Scores

Differences in scoring methods (Method 1 and Method 2) were examined in three ways:

1. One-factor, repeated measures ANOVA, where the factor is the scoring method. The between-subjects F-ratio provide a test of the scoring method effect. By averaging rater scores, the raters were treated as equivalent and were tested for an overall difference.
2. Comparison of average within examinee variance between the two methods. Under a halo effect it was hypothesized that the within-subject variance would be lower for scoring Method 1.
3. Evaluation of within-rater similarity of scores assigned. That is, the average percentage (over examinees) of identical scores assigned to each examinee. If a halo effect were present, we would expect that percentage to be higher under Method 1.

An ANOVA was conducted using a one-factor repeated measures design. The hypothesis of interest was the between-groups effect for Method 1 versus Method 2. The results are presented in Table 1. The between-groups (Method) effects were not significant, implying that there was not an effect due to the method used. The same raters were involved in both studies and method was treated as the only independent variable.

Table 1: One-Factor, Repeated Measures ANOVA

Source	DF	Type II SS	Mean Square	F Value	Prob > F
Method	1	0.4033	0.4033	0.11	0.7350
Error	298	1046.9183	3.5131	—	—

The mean within-subject variance for both Methods 1 and 2 was identical (.26). This finding was consistent with the ANOVA results in showing no difference between the two rating methods.

Table 2 contains the mean percentage of identical scores assigned to each examinee under the different scoring methods. For example, if an examinee received three scores that were identical, the percentage of identical scores would be reported as .75; if an examinee received two scores that were identical, the percentage of identical scores would be reported as .50. The results, presented in Table 2, indicate that Method 1 resulted in a slightly higher percentage of identical scores. However, the differences between methods were insignificant (.715 to .691), additional evidence that there was no difference between methods.

Table 2: Mean Percent Identical Scores Across Domains

Method/Rater	Mean
Method 1/Rater 1	.698
Method 1/Rater 2	.732
Mean Method 1	.715
Method 2/Rater 1	.682
Method 2/Rater 2	.700
Mean Method 2	.691

Additional descriptive statistics were calculated for the two scoring methods and are reported by prompt and by domain in Appendix B. Means by domain were very similar across scoring methods with the largest differences in Domains A and C (see Table B1). Frequency distributions for Method 1 and Method 2 are also provided in Appendix B. Recall that the papers used in the study were selected according to a fixed distribution of scores that followed a bell curve distribution. (This distribution is reflected in the ETS-assigned scores provided in Table B2.) Scoring data show that Method 2 analytic domain scores tended to "clump" in the 2–5 range (see Table B2a–B2d). That is, very few 1's and 6's were assigned in any domain. In Domain A, for example, there were no 5.5 or 6 scores assigned for either Prompt 00043 or 00082, and only one 6 was assigned for prompt 00132 (see Table B2a). The forced distribution of holistic scores tells us that 12 papers that were scored a 6 on the holistic scale were present in this sample but were not recognized by the analytic scoring of Method 2. The same distribution held true for Domain B scores under Method 2 (see Table B2b). No 5.5's or 6's were assigned for Domain C across the three prompts (see Table B2c).

A strong correlation exists between analytic scores from Method 1 and Method 2: the Pearson-r correlation ranged from .83 to .88 across the four domains (see Table B3). Agreement rates between scores from the two scoring methods were high, with perfect agreement ranging from 41% to 52% perfect. The combined adjacent agreement rate for scores between the two methods ranged from approximately 98% to 100% across the domains (see Table B3). The high adjacent agreement rates and strong correlations between scoring methods suggest that there was no significant statistical difference in scoring consistency between the two analytic scoring methods.

Conclusions – Stage 1

The analyses for Stage 1 did not identify an effect from assigning all four analytic scores based on a single read. The results of the ANOVA and comparison of within-subject variances do not indicate significant differences between methods. The descriptive statistics were very similar across methods. Therefore, given the lack of statistical evidence in support of one method over the other, and based on the practical implications associated with implementing each method, Method 1 was the preferred scoring method.

Methodology – Stage 2

Comparing Human and Computerized Scoring with the Analytic Rubric

In the second stage of the study, ACT investigated the ability of both human readers and a computerized Automated Essay Scoring (AES) engine to score Analysis of an Argument (ARG) responses reliably using the analytic domain scales. This ability was measured by (1) the agreement rates (assigning the same scores to each response) between human-to-human scores and AES-to-human scores, as well as by (2) conformity to the expected means and frequency distributions of scores by both human readers and the AES engine, given the existing historical data for both of these measures.

Based on the results from Stage 1, ACT trained six experienced readers (including the original three readers) to score AWA ARG responses using the preferred approach of assigning all four domain scores based on a single reading of an essay response. Three hundred fifty randomly selected responses were selected from historical operational GMAT® data for three prompts (1,050 total responses). Each response was scored independently by two raters. The analytic scores for these responses were the basis for training Vantage Learning's Intellimetric® AES engine to score ARG responses on the four analytic scales.

For each of the three prompts, ACT submitted 300 responses with human scores to Vantage for calibration. Once calibrated, Vantage scored an additional 50 responses per prompt (a total of 150 audit responses) and returned the scores to ACT for analysis and summary. The division of responses between the calibration process and the audit process (300 and 50) was determined after consideration of the relative importance of the training versus the evaluation. The AES engine was optimally trained using responses spread across all scores for all domains. A sample of 300 responses helped to ensure this representation across all four domains.

Data Analysis and Findings – Stage 2

Comparison of Vantage AES Analytic Scoring to ACT Human Analytic Scoring

The Vantage AES engine generated scores in each of the four domains for each of the three prompts, using a total of 150 audit responses.

Table 3 provides a comparison of mean scores. There was a close agreement between the ACT human analytic scores and the Vantage AES analytic scores for each of the domains. (See Tables B4a & B4b in Appendix B for more detail at the prompt level.)

Table 3: Means by Domain for ACT Human and Vantage AES-Generated Scores

Domain	ACT All Prompts	Vantage AES All Prompts
A	3.33	3.28
B	3.19	3.23
C	3.40	3.46
D	3.69	3.71
A+B+C+D	3.40	3.42

This indicates that the Vantage AES engine was able to analytically score essays in a fashion that was very consistent with the analytic scores applied by ACT's human raters. The AES engine showed no bias toward either generally higher or generally lower scores.

Table 4 illustrates that agreement rates were consistently and significantly lower for AES-to-human analytic scores than they were for ACT's human-to-human analytic scores in Domains A (analysis), B (support), and C (organization). Perfect agreement rate for Domain D (language) was consistent between human scores and AES scores. (See Tables B5a–B5d and B6a–B6d in Appendix B for more detail.)

Table 4: Perfect Agreement Rates by Domain for ACT Human and Vantage AES-Generated Scores

Domain	ACT Human/Human	Vantage AES/One Human
A	75.5%	58.0%
B	72.6%	60.7%
C	72.9%	61.3%
D	66.6%	64.0%

Domain A (analysis) showed the most variability in agreement rates between human-to-human and AES-to-human scores, with human-to-human agreement at 75.5% and AES-to-human agreement at 58%. This indicates that, while AES engines may perform well in recognizing the structural linguistics and language features associated with domains B–D, they may be less adept at isolating and evaluating writing features related to logical analysis and critical reasoning.

A comparison of the overall agreement rates (perfect agreement + adjacent agreement) results in more similar results across the two scoring methods. Considering all scores within one point, human-to-human adjacent agreement across the four domains ranged from 98.4% to 99.0%. Vantage AES-to-human adjacent agreement across the four domains similarly ranged from 96.7% to 100.0%. (See Tables B5a–B5d and B6a–B6d in Appendix B for more detail.)

A comparison of the correlations between the ACT-assigned scores and the Vantage AES scores shows a very strong relationship for Domains A, B, and C (see Table 5), indicating that both ACT scoring and Vantage AES scoring were measuring similar characteristics of writing.

Table 5: Correlations by Domain between Human Scores and Vantage AES-Generated Scores

Domain	ACT/Vantage Pearson Correlation
A	.858
B	.867
C	.857
D	.777

Total score frequency distributions (the sum of two scores on a scale that ranges from 1.0 to 6.0 in increments of .5) are very similar when comparing human + human analytic scores and AES + human score (see Tables B8a–B8d and Tables B9a–B9d in Appendix B).

Comparison of Human Analytic Scoring and Existing Holistic Scores

ACT raters provided two independent sets of analytic scores across the four domains for each of the three prompts, for a total of 1,050 responses.

Comparison of mean scores shows close agreement between existing ETS holistic scores and ACT human analytic scores averaged across the four domains [(A+B+C+D)/4]. Table 6 shows the averaged ACT analytic means compared to ETS holistic means for the three ARG prompts.

Table 6: Comparison of Means by Prompt for ACT and ETS Holistic Scores

Prompt	ACT Averaged	ETS Holistic
00043	3.34	3.31
00082	3.43	3.40
00132	3.44	3.32

The ACT analytic domain means most consistent with ETS holistic means were Domain A (analysis) and Domain C (organization). Domain B (support) was consistently rated lower than the holistic mean—3.18 overall, compared to a 3.34 overall holistic mean—and Domain D (language) was consistently rated higher than the holistic mean—3.69 overall, compared to a 3.34 overall holistic mean. (See Tables B4a and B4c in Appendix B for more detail.)

The data indicate that the averaged scores from analytically scored responses were consistent with the holistic scores for these responses. Whereas, under the analytic scoring method, responses were consistently rated lower on the basis of the writing criteria "Supports the critique using relevant supporting reasons and/or examples" (Domain B), and responses were consistently rated higher on the writing criteria of "Demonstrates control of language, including diction, syntax, and conventions of standard written English" (Domain D). These differences appear to offset each other in the overall averaged scores.

A comparison of the correlations between the ACT human analytic scores and the ETS holistic scores shows a very strong relationship between the two scores, especially for Domains A and B (see Table 7).

Table 7: Correlations by Domain between ACT Analytic Scores and ETS Holistic Scores

Domain	ACT Analytic/ETS Holistic Pearson Correlation
A	.863
B	.860
C	.774
D	.727

The strong correlation between the Domain A and B analytics and the ETS holistic argues that these scales are measuring many of the same features of writing. The somewhat more modest, although still high, correlations between the Domain C and D analytics and the ETS holistic argues that these domains are identifying somewhat different features than the holistic rubric.

Comparison of Vantage Analytic Scoring and Existing Holistic Scores

Comparison of mean scores shows close agreement between Vantage analytic scores and existing ETS holistic scores for each of the domains. Table 8 shows the averaged AES analytic means compared to ETS holistic means for the three ARG prompts.

Table 8: Comparison of Means by Prompt for Vantage AES-Generated Scores and ETS Holistic Scores

Prompt	Vantage AES Averaged	ETS Holistic
00043	3.36	3.31
00082	3.46	3.40
00132	3.49	3.32

Vantage AES domain means followed a pattern that was very consistent with ACT human domain means: the domain means most consistent with ETS holistic means were Domain A (analysis) and Domain C (organization). Domain B (support) was consistently rated lower than the holistic mean—3.23 overall, compared to a 3.34 overall holistic mean—and Domain D (language) was consistently rated higher than the holistic mean—3.71 overall, compared to a 3.34 overall holistic mean. (See Tables B4b and B4c in Appendix B for more detail.)

The frequency distributions of both human and AES analytic scores identified a clumping of scores around the middle of the score scale (3 and 4 score points), slightly greater than that found with the ETS holistic scale. (Refer to Table B7 for holistic score frequency distributions, and see Tables B8a–B8d and B9a–B9d in Appendix B for analytic distributions.)

The frequency distributions of both human and AES analytic scores also showed a smaller percentage of upper-range analytic scores (5 and 6 scorepoints) compared to the distribution of the holistic scores. Human analytic scores of 5 and 6 accounted for approximately 8% to 10% of overall scores given, and AES analytic scores of 5 and 6 accounted for approximately 4% to 6% of overall scores given, whereas ETS holistic scores of 5 and 6 accounted for 20% of scores. (See Tables B7, B8a–B8d, and B9a–B9d in Appendix B.) According to this data, both human and AES analytic scoring showed the same trend toward a lower percentage of upper-range scores.

During early ACT pilot studies of training materials, comparisons made between ACT human holistic scores and existing ETS holistic scores, as well as comparisons between AES holistic scores and existing ETS holistic scores, showed a similar pattern of difference in frequency distribution, with a tendency by ACT and Vantage AES to assign a lower percentage of upper-range scores.

Conclusions – Stage 2

This study's analytic scoring results demonstrate ACT's human raters' and the Vantage AES's ability to score responses on the four analytic domains accurately. At the same time, the means for the analytic domains revealed variation on Domains B and D that had not been apparent under the holistic scoring system, indicating that additional information would be provided with the analytic rubrics that is not provided with the holistic rubric.

It does appear, however, that the information provided by the analytic scores would indeed offer more information to admissions officers about each examinee's varying levels of ability on the four key writing criteria of the ARG analytic scoring rubric. Admissions officers could have greater flexibility by deciding which criteria to weigh more heavily in order to meet their individual institutional needs, or could alternatively use a scale that sums the scores.

The closeness of ACT summed means and ETS holistic means in this study indicates that such summed scores would still provide a baseline of measurement that is consistent with the holistic scores assigned in past administrations, even while the variation of individual analytic scores assigned on each response indicates that institutions could achieve different results by weighting some domains more than others in their admissions decisions.

One could also infer that, as a cohort, AWA examinees demonstrate in their responses stronger language skills (as defined by the rubric) than skills in supporting ideas with reasoning or examples, a difference that was not apparent when looking only at holistic scores.

Differences between ACT pilot study scoring and ETS operational scoring may have accounted for some of the trend toward lower percentages of scores in the 5 to 6 range, both when scoring holistically during earlier pilot studies and when scoring with the analytic rubric in the current study. Considering that ACT's subsequent operational holistic scoring distributions (which include both ACT human and Vantage AES scores) have not shown a lack of upper range scores or any lack of "score spread" across the score scale, this finding is likely not a cause for concern for future operational scoring using an analytic rubric. The trend has not been seen in operational holistic scoring, which grew out of the earlier pilot studies and which has involved training Vantage AES using existing ETS holistically scored responses. However, some "clumping" effect around the 3 and 4 score points, due to analytic domain scoring, may be suggested by these data.

Analytic scoring would thus appear to provide more detailed information about the examinee's writing ability. However, this more-specific information only relates to analytical writing ability as measured by the Arguments rubric. Since the Arguments task only involves the more evidence-based analysis of an existing body of prompt information, the scope of this more-detailed information about writing ability is limited. The kind of information about analytical writing ability that is measured by the Issues task, including the generative analytical writing abilities, language control, and use of traditional essay structure that are required for success on the Issues task in particular, is not measured by the analytic rubric developed for the Argument task.

Next Steps

The results of the study indicate that the AWA responses can be reliably scored by trained human readers and by the Vantage AES. However, before implementing the analytic rubric on an operational basis, the following next steps may be considered.

1. Survey users of the AWA scores to verify that the analytic information is consistent with their needs. Guiding questions may include:
   - Does the analytic rubric provide valuable information that helps to inform the admission decision?
   - Does the use of the analytic rubric provide the appropriate balance of language skills (Domain D) and argument skills (Domains A–C)?
2. Examine the relationship between the AWA analytic rubric scores and other GMAT® scores (Verbal and Quantitative) to verify that this relationship is consistent with the goals of the GMAT® exam.
3. Continue to refine the calibration of the AES to ensure that the scoring engine can consistently identify scores at the upper end of each domain (5- and 6-level responses). A focused study that identifies certain types of responses that receive high scores in specific domains may help this refinement.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC® Research and Development department at research@gmac.com.

Acknowledgements

This document was prepared by Tim Burden, Catherine Welch, and John C. Kalohn of ACT, Inc. The study was facilitated by Jay Berry, Rosanne Cook, and Dan Shaw of ACT, Inc.

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the Graduate Management Admission Council®.

Appendix A

GMAT® Analytic Scoring Rubric for Analysis of an Argument

Score	Domain A: *Identifies and analyzes significant flaws in the argument.*	Domain B: *Supports the critique using relevant supporting reasons and/or examples.*	Domain C: *Develops a clearly organized and coherent response.*	Domain D: *Demonstrates control of language, including diction, syntax, and conventions of standard written English.*
1	Demonstrates little or no ability to understand or analyze the argument.	Demonstrates little or no ability to support ideas.	Provides little or no evidence of the ability to develop an organized response.	Has severe and persistent errors in language use that result in incoherence.
2	Does not present a critique based on logical analysis of the argument, or may not identify any significant flaws in the argument.	Provides little, if any, relevant support.	Provides a little evidence of the ability to develop an organized response, but response is generally disorganized.	Has serious and frequent errors in language use that interfere with the communication of meaning.
3	Some analysis of the argument is present, but either fails to identify or fails to analyze most of the significant flaws in the argument; may analyze tangential or irrelevant matters; or may reason poorly.	Provides support of limited relevance and value for the main points of the critique.	Demonstrates limited ability to develop a logically organized response.	Demonstrates some control of language but does not clearly convey ideas, due to errors in language use that interfere with the communication of meaning.
4	Identifies significant flaws in the argument and analyzes them adequately.	Adequately supports the main points of the critique.	Develops an adequately organized response but uses simple transitions (if any) to connect ideas.	Demonstrates enough control of language to convey ideas with sufficient clarity, and also may have some flaws in conventions.
5	Clearly identifies significant flaws in the argument and analyzes them in a generally thoughtful way.	Effectively supports the main points of the critique, and may show some persuasiveness.	Develops a clear and logically organized response, and connects ideas with appropriate transitions.	Demonstrates control of language, including diction, some syntactic variety, and facility with conventions.
6	Clearly identifies significant flaws in the argument and analyzes them insightfully.	Persuasively supports the main points of the critique.	Develops a focused and logically organized response, and connects ideas with clear and effective transitions.	Demonstrates strong control of language, including precise diction, effective syntactic variety, and facility with conventions.

Appendix B

Table B1: Stage 1 Means and Standard Deviation for Analytic Scores, by Method

Table B1: Means (and Standard Deviation) for ETS Holistic and Analytic Method 1*

Domain	Prompt 00043		Prompt 00082		Prompt 00132		Total Across Prompts	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
A	3.40	3.32	3.55	3.43	3.37	3.37	3.44 (1.20)	3.37 (1.00)
B	3.22	3.21	3.37	3.18	3.18	3.21	3.26 (1.23)	3.20 (0.96)
C	3.24	3.21	3.43	3.44	3.32	3.56	3.33 (1.01)	3.40 (0.91)
D	3.68	3.69	3.91	3.67	3.63	3.70	3.74 (0.96)	3.69 (0.90)
A+B+C+D/4	3.39	3.36	3.57	3.43	3.38	3.46	3.44 (1.08)	3.42 (0.83)

*The ETS holistic mean and (standard deviation) for this sample were 3.80 (1.40)

Table B2: Stage 1 Frequency Distributions: Analytic Method 1 vs. Method 2

Table B2a: Domain A Analytic Method 1 vs. Method 2, with ETS Holistic

Score	Prompt 00043		Prompt 00082		Prompt 00132		ETS Holistic per Prompt
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	
1	2	0	1	1	2	1	4
1.5	0	2	1	1	3	0	—
2	7	5	7	6	7	7	9
2.5	4	3	3	3	5	4	—
3	11	10	12	12	9	14	12
3.5	5	10	4	4	3	6	—
4	10	12	8	12	7	7	12
4.5	6	5	3	7	3	7	—
5	3	2	7	4	9	3	9
5.5	1	0	2	0	1	0	—
6	1	0	2	0	1	1	4

Table B2b: Domain B Analytic Method 1 vs. Method 2, with ETS Holistic

Score	Prompt 00043		Prompt 00082		Prompt 00132		ETS Holistic per Prompt
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	
1	2	1	3	1	3	1	4
1.5	1	0	1	1	6	1	—
2	10	8	8	8	5	11	9
2.5	3	7	3	5	5	3	—
3	9	8	10	12	9	11	12
3.5	6	13	5	9	3	6	—
4	13	6	9	8	7	9	12
4.5	2	5	1	3	5	5	—
5	3	2	7	3	5	2	9
5.5	0	0	1	0	1	0	—
6	1	0	2	0	1	1	4

Table B2c: Domain C Analytic Method 1 vs. Method 2, with ETS Holistic

Score	Prompt 00043		Prompt 00082		Prompt 00132		ETS Holistic per Prompt
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	
1	1	1	1	2	2	2	4
1.5	2	—	1	—	1	1	—
2	6	6	7	5	5	4	9
2.5	4	8	5	3	6	3	—
3	10	11	5	10	10	10	12
3.5	10	8	9	4	4	9	—
4	14	12	14	19	16	16	12
4.5	2	3	4	5	3	2	—
5	0	1	2	2	1	3	9
5.5	0	0	0	0	2	0	—
6	1	0	2	0	0	0	4

Table B2d: Domain D Analytic Method 1 vs. Method 2, with ETS Holistic							
	Prompt 00043		Prompt 00082		Prompt 00132		ETS Holistic per Prompt
Score	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	
1	0	0	0	0	1	1	4
1.5	0	1	0	1	1	—	—
2	1	1	1	3	4	3	9
2.5	10	6	5	5	6	2	—
3	5	5	8	8	5	11	12
3.5	4	9	3	5	7	5	—
4	22	16	21	17	13	16	12
4.5	3	9	2	4	3	5	—
5	3	3	5	6	9	6	9
5.5	1	0	3	0	0	0	—
6	1	0	2	1	1	1	4

Table B3: Stage 1 Agreement between Scores: Analytic Method vs. Method 2

Table B3: Stage 1 Agreement between Scores – Analytic Method vs. Method 2				
	Domain			
	A	B	C	D
Percent Perfect Agreement	51.3%	41.3%	48.7%	52.0%
Combined Percent Adjacent Agreement	99.3%	98.6%	100%	99.3%
Percent Requiring Resolution	.7%	1.4%	0.0%	0.7%
Inter-rater Correlation	.88	.88	.88	.83

Table B4: Means for ACT and Vantage AES Analytic Scoring and ETS Holistic Scoring

Table B4a: Means (and Standard Deviation) for ACT Analytic Scoring				
	ACT			
Domain	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts (SD)
A	3.21	3.39	3.40	3.33 (1.12)
B	3.03	3.27	3.27	3.18 (1.13)
C	3.32	3.45	3.44	3.40 (0.98)
D	3.65	3.72	3.70	3.69 (0.86)
A+B+C+D	3.34	3.43	3.44	3.40 (0.97)

Table B4b: Means (and Standard Deviation) for Vantage AES Analytic Scoring

Domain	Vantage AES			
	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts (SD)
A	3.28	3.36	3.41	3.28 (1.03)
B	3.08	3.30	3.31	3.23 (1.05)
C	3.39	3.49	3.49	3.46 (0.92)
D	3.69	3.70	3.75	3.71 (0.75)
A+B+C+D	3.36	3.46	3.49	3.42 (0.91)

Table B4c: Table B4a: Means (and Standard Deviation) for ETS Holistic Scoring

	ETS Holistic			
	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts (SD)
Holistic	3.31	3.40	3.32	3.34 (1.30)

Table B5: Agreement Rates, by Domain and by Prompt, for ACT Analytic Scoring

Table B5a: Domain A Agreement Rates for ACT Analytic Scoring by Prompt

Domain A	ACT Human/Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	79.7%	77.8%	69.3%	75.5%
Adjacent	18.9%	21.8%	29.2%	23.3%
Discrepant	1.1%	0.6%	1.4%	1.1%

Table B5b: Domain B Agreement Rates for ACT Analytic Scoring by Prompt

Domain B	ACT Human/Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	74.2%	75.6%	68.2%	72.6%
Adjacent	25.2%	23.3%	30.1%	26.2%
Discrepant	0.3%	1.2%	1.7%	1.1%

Table B5c: Domain C Agreement Rates for ACT Analytic Scoring by Prompt				
Domain C	ACT Human/Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	71.9%	79.4%	67.6%	72.9%
Adjacent	27.2%	19.2%	31.2%	25.9%
Discrepant	0.9%	1.5%	1.2%	1.1%

Table B5d: Domain D Agreement Rates for ACT Analytic Scoring by Prompt				
Domain D	ACT Human/Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	72.8%	68.6%	58.5%	66.6%
Adjacent	25.5%	29.1%	39.5%	31.4%
Discrepant	1.4%	1.7%	1.7%	1.6%

Table B6: Agreements Rates, by Domain and by Prompt, for Vantage AES-Generated Scores

Table B6a: Domain A Agreement Rates for Vantage AES-Generated Scores				
Domain A	Vantage AES/One Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	52.0%	52.0%	70.0%	58.0%
Adjacent	46.0%	40.0%	30.0%	38.7%
Discrepant	2.0%	8.0%	0.0%	3.3%

Table B6b: Domain B Agreement Rates for Vantage AES-Generated Scores				
Domain B	Vantage AES/One Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	60.0%	66.0%	56.0%	60.7%
Adjacent	36.0%	34.0%	42.0%	37.3%
Discrepant	4.0%	0.0%	2.0%	2.0%

Table B6c: Domain C Agreement Rates for Vantage AES-Generated Scores

Domain C	Vantage AES/One Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	60.0%	62.0%	62.0%	61.3%
Adjacent	40.0%	38.0%	38.0%	38.7%
Discrepant				

Table B6d: Domain D Agreement Rates for Vantage AES-Generated Scores

Domain D	Vantage AES/One Human			
Agreement	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
Perfect	60.0%	66.0%	66.0%	64.0%
Adjacent	36.0%	32.0%	32.0%	33.3%
Discrepant	4.0%	2.0%	2.0%	2.7%

Table B7: ETS Holistic Frequency Distribution

Table B7: ETS Holistic Frequency Distribution

Score	Percent Score Point
1	9.1%
2	20.2%
3	24.5%
4	26.0%
5	13.7%
6	6.5%

Table B8: Frequency Distributions for ACT Analytic Scoring

Table B8a: ACT Raters (Human+Human)/2, or 3rd Rating for Domain A

Domain A	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	4.3%	5.5%	5.7%	5.2%
1.5	2.0%	0.9%	2.3%	1.7%
2.0	18.1%	13.1%	8.9%	13.3%
2.5	4.6%	6.4%	7.4%	6.1%
3.0	24.4%	19.8%	23.5%	22.6%
3.5	8.0%	8.4%	9.5%	8.6%
4.0	27.8%	28.8%	22.6%	26.1%
4.5	3.7%	4.4%	9.2%	5.8%
5.0	5.2%	9.0%	6.3%	6.8%
5.5	0.9%	1.7%	0.9%	1.2%
6.0	1.1%	2.0%	3.7%	2.3%
Total	100.0%	100.0%	100.0%	100.0%

Table B8b: ACT Raters (Human+Human)/2, or 3rd Rating for Domain B

Domain B	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	6.3%	6.1%	6.6%	6.3%
1.5	3.7%	2.0%	2.3%	2.7%
2.0	18.6%	13.7%	12.0%	14.8%
2.5	7.2%	6.4%	8.9%	7.5%
3.0	25.2%	23.0%	22.3%	23.5%
3.5	8.6%	9.3%	11.5%	9.8%
4.0	20.3%	25.0%	19.2%	21.5%
4.5	4.9%	4.1%	6.9%	5.3%
5.0	3.2%	7.3%	6.0%	5.5%
5.5	0.9%	1.5%	0.9%	1.1%
6.0	1.1%	1.7%	3.4%	2.1%
Total	100.0%	100.0%	100.0%	100.0%

Table B8c: ACT Raters (Human+Human)/2, or 3rd Rating for Domain C				
Domain C	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	2.9%	3.2%	5.4%	3.8%
1.5	2.0%	1.2%	3.2%	2.1%
2.0	9.7%	8.4%	7.4%	8.5%
2.5	6.6%	5.5%	5.4%	5.9%
3.0	23.2%	23.0%	18.3%	21.5%
3.5	14.3%	8.4%	12.9%	11.9%
4.0	33.8%	40.4%	31.2%	36.1%
4.5	4.0%	3.5%	7.4%	5.0%
5.0	2.9%	4.1%	4.0%	3.6%
5.5	0.6%	0.9%	2.6%	1.3%
6.0	0.0%	1.5%	2.0%	1.2%
Total	100.0%	100.0%	100.0%	100.0%

Table B8d: ACT Raters (Human+Human)/2, or 3rd Rating for Domain D				
Domain D	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	0.8%	0.6%	0.3%	0.5%
1.5	0.9%	2.0%	2.3%	1.7%
2.0	5.4%	3.2%	3.7%	4.1%
2.5	5.4%	4.1%	7.4%	5.7%
3.0	18.1%	14.0%	18.6%	16.9%
3.5	11.5%	16.9%	15.8%	14.7%
4.0	43.8%	44.8%	28.4%	39.0%
4.5	6.9%	4.4%	11.7%	7.7%
5.0	5.7%	6.1%	6.6%	6.1%
5.5	1.1%	1.7%	2.3%	1.7%
6.0	0.6%	1.7%	2.6%	1.6%
Total	100.0%	100.0%	100.0%	100.0%

Table B9: Frequency Distributions for Vantage AES-Generated Scores

Table B9a: Vantage: (AES+One Human)2 for Domain A

Domain A	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	0.0%	4.0%	2.0%	2.0%
1.5	8.0%	2.0%	4.0%	4.7%
2.0	10.0%	6.0%	10.0%	8.7%
2.5	10.0%	6.0%	10.0%	8.7%
3.0	20.0%	26.0%	26.0%	21.0%
3.5	16.0%	20.0%	6.0%	14.0%
4.0	22.0%	20.0%	26.0%	22.7%
4.5	12.0%	10.0%	6.0%	9.3%
5.0	2.0%	4.0%	6.0%	4.0%
5.5	0.0%	2.0%	4.0%	2.0%
6.0	0.0%	0.0%	0.0%	0.0%
Total	100.0%	100.0%	100.0%	100.0%

Table B9b: Vantage: (AES+One Human)2 for Domain B

Domain B	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	6.0%	4.0%	4.0%	4.7%
1.5	6.0%	2.0%	2.0%	3.3%
2.0	14.0%	10.0%	6.0%	10.0%
2.5	6.0%	14.0%	18.0%	12.7%
3.0	26.0%	20.0%	24.0%	23.3%
3.5	16.0%	12.0%	10.0%	13.3%
4.0	6.0%	24.0%	20.0%	20.0%
4.5	6.0%	4.0%	8.0%	6.0%
5.0	0.0%	8.0%	4.0%	4.0%
5.5	2.0%	2.0%	4.0%	2.7%
6.0	0.0%	0.0%	0.0%	0.0%
Total	100.0%	100.0%	100.0%	100.0%

Table B9c: Vantage: (AES+One Human)2 for Domain C				
Domain C	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	4.0%	2.0%	4.0%	3.3%
1.5	0.0%	2.0%	8.0%	3.3%
2.0	6.0%	2.0%	2.0%	3.3%
2.5	12.0%	8.0%	8.0%	9.3%
3.0	18.0%	20.0%	14.0%	17.3%
3.5	22.0%	20.0%	16.0%	19.3%
4.0	32.0%	36.0%	38.0%	35.3%
4.5	6.0%	6.0%	6.0%	6.0%
5.0	0.0%	2.0%	2.0%	1.3%
5.5	0.0%	2.0%	0.0%	0.7%
6.0	0.0%	0.0%	2.0%	0.7%
Total	100.0%	100.0%	100.0%	100.0%

Table B9d: Vantage: (AES+One Human)2 for Domain D				
Domain D	Prompt 00043	Prompt 00082	Prompt 00132	Across Prompts
1.0	0.0%	0.0%	0.0%	0.0%
1.5	0.0%	2.0%	0.0%	0.7%
2.0	0.0%	0.0%	4.0%	1.3%
2.5	8.0%	10.0%	6.0%	8.0%
3.0	22.0%	20.0%	18.0%	20.0%
3.5	22.0%	10.0%	16.0%	16.0%
4.0	36.0%	42.0%	36.0%	38.0%
4.5	6.0%	10.0%	10.0%	8.7%
5.0	6.0%	4.0%	8.0%	6.0%
5.5	0.0%	0.0%	0.0%	0.0%
6.0	0.0%	2.0%	2.0%	1.3%
Total	100.0%	100.0%	100.0%	100.0%