J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 4, Number 4 · March 2006

An Evaluation of the IntelliMetric℠ Essay Scoring System

Lawrence M. Rudner, Veronica Garcia, & Catherine Welch

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 4, Number 4

An Evaluation of the IntelliMetric℠ Essay Scoring System

Lawrence M. Rudner, Veronica Garcia, & Catherine Welch

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

This article has been peer reviewed and printed with permission from the Graduate Management Admission Council® (GMAC®) to the Journal of Technology, Learning, and Assessment (JTLA).

Preferred citation:

Rudner, L. M., Garcia, V., & Welch, C. (2006). An Evaluation of the IntelliMetric℠ Essay Scoring System. *Journal of Technology, Learning, and Assessment, 4*(4). Available from http://www.jtla.org

Abstract:

This report provides a two-part evaluation of the IntelliMetric℠ automated essay scoring system based on its performance scoring essays from the Analytic Writing Assessment of the Graduate Management Admission Test™ (GMAT™). The IntelliMetric system performance is first compared to that of individual human raters, a Bayesian system employing simple word counts, and a weighted probability model using more than 750 responses to each of six prompts. The second, larger evaluation compares the IntelliMetric system ratings to those of human raters using approximately 500 responses to each of 101 prompts. Results from both evaluations suggest the IntelliMetric system is a consistent, reliable system for scoring AWA essays with a perfect + adjacent agreement on 96% to 98% and 92% to 100% of instances in evaluations 1 and 2, respectively. The Person r correlations of agreement between human raters and the IntelliMetric system averaged .83 in both evaluations.

J·T·L·A

An Evaluation of the IntelleMetric℠ Essay Scoring System

Lawrence M. Rudner & Veronica Garcia
Graduate Management Admission Council
Catherine Welch
Assessment Innovations at ACT, Inc.

Introduction

The Graduate Management Admission Council® (GMAC®) has long benefited from advances in automated essay scoring. When GMAC adopted ETS® e-rater® in 1999, the Council's flagship product, the Graduate Management Admission Test® (GMAT®), became the first large-scale assessment to incorporate automated essay scoring. The change was controversial at the time (Iowa State Daily, 1999; Calfee, 2000). Though some may still find it controversial, automated essay scoring is now widely accepted as a tool to compliment, but not replace, expert human raters.

Starting in January 2006, ACT, Inc. will be responsible for GMAT test development and scoring, and a new automated essay scoring system will be utilized in conjunction with the ACT™ contract. ACT included the IntelliMetric Essay Scoring System of Vantage Learning as part of their initial proposal. Before approving the Vantage subcontract, GMAC wanted assurance that the IntelliMetric (IM) system could reasonably approximate the scores provided by human raters on the GMAT Analytic Writing Assessment.

This paper provides an overview of the GMAT Analytic Writing Assessment and part of the results of an evaluation of the IntelliMetric system. The evaluation is twofold. An initial evaluation examines the performance of the IntelliMetric system based on a sample of responses to six essays. Results for the IntelliMetric system are compared to individual human raters, a primitive Bayesian system using simple word counts, and a weighted probability model. A second evaluation is based on the comprehensive system reliability demonstration presented by Vantage to both ACT and GMAC. This second evaluation relies solely on comparisons to scores calculated by human raters, as such agreement will be the prime measure of performance during operational use of the IntelliMetric system in 2006.

Background

The GMAT Analytic Writing Assessment

The Analytical Writing Assessment (AWA) is designed as a direct measure of the test taker's ability to think critically and communicate ideas. The AWA consists of two 30-minute writing tasks—Analysis of an Issue and Analysis of an Argument.

For Analysis of an Issue prompts, the examinee must analyze a given issue or opinion and explain their point of view on the subject by citing relevant reasons and/or examples drawn from experience, observations, or reading.

For Analysis of an Argument prompts, the examinee must read a brief argument, analyze the reasoning behind it, and then write a critique of the argument. In this task, the examinee is not asked to state her opinion, but to analyze the one given. The examinee may, for example, consider what questionable assumptions underlie the thinking, what alternative explanations or counterexamples might weaken the conclusion, or what sort of evidence could help strengthen or refute the argument.

For both tasks, the examinee writes her response on the screen using rudimentary word-processing functions built into the GMAT test driver software. Scratch paper or erasable noteboards are provided at the test center for use by examinees in planning their responses. Because there is no one right answer, all GMAT prompts are available on-line for candidates to review prior to taking the test.

Prompts are initially scored by two human raters following detailed scoring rubrics. If the two reviewers differ by more than one score point on a 0 to 6 point scale, a third reader adjudicates scores. Once a sufficient number of responses to a given prompt have been hand-scored, an automated essay scoring model is developed and evaluated for the prompt. If an acceptable model can be formulated, the automated system replaces one of the two human raters. The automated essay scoring system can be viewed as an amalgamation of all the human raters who have scored the item, and use of automated essay scoring can be viewed as a check on the human rater.

Studies of the reliability and consistency of AWA prompt scoring by either human raters or automated systems raise the related issue of the validity of the AWA prompts themselves in predicting viable candidacy for graduate management education, one of the original goals in adding the AWA section to the GMAT exam. Perhaps not unexpectedly, studies conducted through the Validity Study Service at GMAC have found that, as an individual element, AWA scores tend to be the least predictive GMAT

score. Although there are many programs where GMAT AWA out predicts GMAT Quant, a summary of validity data from 277 studies conducted from 1997-2004 found a mean predictive validity value for the AWA score of .184 with an interquartile range of .101 to .277. In contrast, the mean validity coefficients for Verbal and Quantitative scores are .323 and .331, respectively (Talento-Miller & Rudner, 2005). However, when the AWA scores are used in combination with Verbal, Quantitative, and Undergraduate Grade Point Average, the mean predictive validity is an impressive .513.

Essay Scoring Approaches

Interest and acceptance of automated essay scoring appears to be growing, as is evident in the increasing number of references in the academic media over the last few years. In January 2005, one on-line bibliography contained 175 references to machine scoring (Haswell, 2005). A recent book by Shermis and Burstein (2003), the first to focus entirely on automated essay scoring and evaluation, provides descriptions of all the major approaches (see reviews by Rudner, 2004; Myford, 2004). An on-line summary of several approaches is also provided by Valenti, Neri and Cucchiarelli (2003).

Despite the number of approaches available, the basic procedure is the same. A relatively large set of pre-scored essays responding to one prompt are used to develop or calibrate a scoring model for that prompt. Once calibrated, the model is applied as a scoring tool. Models are then typically validated by applying them to a second, but independent, set of pre-scored items.

The following is a description of the three approaches used in this paper. The first approach, that of the IntelliMetric system, is a true automated scoring system. The other two provide a basis for comparison in the initial evaluation of the IntelliMetric system. The Bayesian approach used in this evaluation employs only simple word counts in building a model. The probability approach is simple random draws from the AWA score distribution, which provides a comparison with chance. All three evaluations compare the IntelliMetric system with scores generated from human raters, which will be the measure of performance during operational use.

The IntelliMetric System

Since first developing the IntelliMetric essay scoring engine in 1998, Vantage Learning has applied their patented technology to become one of the lead providers of writing instruction and automated essay scoring service. Vantage's online, portfolio-based writing instruction program, MY Access!™, which is based on the IntelliMetric system, is widely used in

classrooms and has won numerous awards for innovation, including two Codie® awards in 2005 and finalist nominations for the previous two years. The automated essay system is used in several states, including California, Pennsylvania, Massachusetts, Virginia, Oregon, and Texas; as well as several major testing companies, including The College Board, ACT, Harcourt Assessment, Inc., CTB/McGraw Hill, and Thomson Learning. Microsoft®, Apple® Computer, AOL®, and Sun Microsystems® also license Vantage technology.

Vantage's corporate strategy is to protect the IntelliMetric system, one of their primary intellectual assets, by treating details of the technology as a proprietary trade secret. The chapter by Vantage in the Shermis and Burstein (2003) book describes only the general concepts of the technology behind their product.

While the IntelliMetric system continues to be protected by various patents and many details remain trade secrets, a recent paper by Elliot and Mikulas (2004) provides a great deal of insight into the logic of the IntelliMetric system. Text is parsed to flag the syntactic and grammatical structure of the essay. Each sentence is tagged with regard to parts of speech, vocabulary, sentence structure, and concept expression. Several patented technologies are then applied to examine the text using a variety of techniques, including morphological analysis, spelling recognition, collocation grammar, and word boundary detection. A 500,000 unique word vocabulary and 16 million word concept net are employed at this stage of the analysis. More than 500 linguistic and grammatical features are tagged.

After the tagging, the data is coded to support computation of multiple mathematical models. Each model associates features extracted from the text with the scores assigned in the training set. The models differ in their mathematical form and with respect to the included variables. The IntelliMetric system employs a proprietary optimization technique to integrate the information from the different models to yield a single assigned score. Vantage views the use of multiple mathematical models as analogous to using multiple judges.

Simple Word Counts

There is a collection of well-developed literature and several commercial applications using Bayes Theorem in text classification (c.f. Mitchell, 1997). Applied to essay scoring, the concept is to identify the words or phrases most closely associated with each essay score. The training essays are used to compute the conditional probabilities of each word being associated with each score group. Applied to text classification, calibration data sets typically have at least 1,000 cases.

Each new essay is evaluated as the product of the probabilities of the presence or absence of each calibrated word in the essay. The score category with the highest posteriori probability is assigned to the new essay. Rudner and Liang (2001) found that with the right combination of options and a large calibration data set, this approach was able to correctly classify 80% of the tested essays into one of two groups.

The model was applied using the public domain software BETSY—Bayesian Essay Test Scoring sYstem, available at http://edres.org/Betsy. Although BETSY provides a range of options, only simple word counts were used in this examination.

Probabilistic Modeling

An evaluation of the GMAT AWA prompts by ACT (2004) found that the distribution of scores were almost identical for each prompt with 87% of the candidates obtaining scores of 3, 4, or 5. The probabilistic model approach assigned scores to each essay by randomly drawing a score from the AWA frequency distribution used to provide the most recent norming information.

Investigation 1: Initial Analysis on Six AWA Prompts

Method

A sample of essays responding to three Analysis of an Argument prompts and three Analysis of an Issue prompts formed the basis for this analysis. For each prompt, approximately 270 essays were tagged as training essays and 500 essays were tagged for the validation sets.

In order to test the ability of the scoring software to detect common "cheating" techniques, 13 essays were fabricated and added to the 500 responses in the validation sets for each prompt. Five essays were off-topic and written in response to a different prompt of the same type (*Issues or Arguments*). Five essays were off-topic and written in response to a different prompt of a different type. One essay was a simple repetition of the entire prompt. Another essay consisted of multiply repeated text, and the final fabricated essay was approximately half a genuine response and half a repetition of the prompt.

The fabricated response essays were randomly inserted into the validation files among the other 500 essays and randomly assigned IDs that were consistent with the IDs of the surrounding essays.

The essays were then sent to Vantage for scoring. The transmitted CD-ROM contained two folders for each prompt—one named *Training* and

the other named *Validation*. The files within the Training folders contained approximately 270 essays along with two ratings per essay. The essays in the Validation folder had no ratings. The task for Vantage was to model each of the prompts using the essays in the Training folder, blindly apply the models to each of the essays in the Validation folder, and then send computer assigned scores for each.

Score ratings provided by Vantage's IntelliMetric system (IM) were compared to the original reader-assigned ratings, the Bayesian system based on word frequency (COUNTS), and weighted random numbers representing chance accuracy (PROB).

Two human ratings, and occasionally a third adjudicated rating, were available in the original data. The human rating used for comparison to the automated essay score was selected randomly from Rater 1 and Rater 2, unless there was a third rater, in which case the third rating was used.

Analyses for each prompt include the following:

- Agreement Statistics (cross tables of ratings assigned, perfect, adjacent, discrepant, and perfect + adjacent agreement counts and percentages)
- Descriptive Statistics (rating means and standard deviations)
- Correlation Analysis (Pearson correlation coefficients)

For a baseline, a similar analysis was done for the original human rating 1 versus human rating 2, COUNTS, and PROB. In addition, a check was done on whether each of the fabricated essays was flagged appropriately.

Results

Scoring

Summary results of the scoring by IM, COUNTS, and PROB using the 500 validation essays for each prompt are shown in Tables 1 to 6. Table 1 provides a summary of the agreement statistics across the six GMAT prompts (three Arguments and three Issues) for each of the scoring models. Table 2 provides a comparison of mean scores for Rater 1 and Rater 2 from the original baseline data. Tables 3 to 6 provide a comparison of mean scores from each of the automated scoring engines to the original scores. Effect sizes are also included in Tables 2 to 6.

Table 1 indicated that the percent of perfect + adjacent agreement ranged from .94 to .98 over the six prompts using the original Rater 1 and Rater 2 data. The average Pearson correlation over the six prompts was .830. Rater 1 and Rater 2 had an average discrepancy of .041 across the six prompts, resulting in slightly more than 4% of the scores needing adjudication.

For the six GMAT prompts, the perfect + adjacent agreement, perfect agreement, and Pearson correlations for the IntelliMetric system were extremely close, and occasionally better, than the corresponding values for two human readers. The values of these statistics were also much higher for IM than they were for simple word counts and weighted random draws. The percent of perfect + adjacent agreement of IM ranged from .96 to .98 over the six prompts, with a slightly higher average than two human readers. The Pearson correlation calculated using IM differed from the Pearson correlation calculated using Original scores by no more than .03 for any single prompt. The IM average Pearson correlation over the six prompts was .833, the same as the Original average. IM had an average discrepancy (more than 2 point difference) of .032 across the six prompts. This would result in a little more than 3% of the scores needing adjudication

Table 1: Summary of Agreement Statistics with the Original Reader Scores for Six GMAT Prompts

Prompt	Content	Prompt IDs Comparison	Original Reader	Vantage IM	COUNTS	PROB
1	Argument	Perfect	.56	.54	.31	.28
		Perf + Adj	.94	.96	.73	.71
		Pearson	.79	.80	.53	-.02
2	Argument	Perfect	.54	.54	.31	.27
		Perf + Adj	.95	.98	.75	.72
		Pearson	.81	.84	.51	-.04
3	Argument	Perfect	.56	.54	.29	.27
		Perf + Adj	.96	.96	.72	.72
		Pearson	.83	.82	.39	.03
4	Issue	Perfect	.62	.62	.24	.28
		Perf + Adj	.96	.98	.70	.73
		Pearson	.85	.87	.41	.03
5	Issue	Perfect	.59	.60	.31	.27
		Perf + Adj	.98	.98	.73	.71
		Pearson	.85	.84	.43	-.04
6	Issue	Perfect	.59	.55	.30	.29
		Perf + Adj	.97	.96	.77	.74
		Pearson	.85	.83	.48	.07

Perf + Adj = Perfect + Adjacent Agreement
Original Reader: Agreement between Original Rater 1 and Rater 2
Vantage IM: Agreement of Vantage's IntelliMetric system with Original Readers
COUNTS: Agreement of Bayesian system based on word frequency with Original Readers
PROB: Agreement of weighted random number (chance accuracy)

Table 2 provides a baseline comparison with two human readers. Rater 1 mean scores were not meaningfully different compared to Rater 2 mean scores for each of the six essays. Effect sizes ranged from .01 to .06.

Table 2: Rater 1 Mean Scores Compared to Rater 2 Mean Scores

	Rater 1		Rater 2		
Prompt	Mean	S.D.	Mean	S.D.	Effect Size
1	3.54	1.26	3.56	1.29	.02
2	3.55	1.29	3.49	1.25	.05
3	3.57	1.24	3.53	1.30	.03
4	3.55	1.29	3.60	1.28	.04
5	3.55	1.26	3.56	1.25	.01
6	3.59	1.24	3.51	1.30	.06

The mean scores provided by IM, as shown in Table 3, were slightly higher than the Original scores for each of the six essays. While the effect sizes are small, ranging from .08 to .15, the fact that the IM means were higher for all six essays raises the possibility that their might be a slight upward bias, albeit minimal, with IM scores. (Investigation 2 of this paper presents data with positive and negative effect sizes, reducing concern in this area.)

Table 3: Vantage Mean Scores Compared to Original Reader Mean Scores

	Vantage		Original Reader		
Prompt	Mean	S.D.	Mean	S.D.	Effect Size
1	3.70	1.51	3.53	1.58	.11
2	3.66	1.59	3.53	1.60	.08
3	3.77	1.50	3.54	1.62	.15
4	3.73	1.57	3.57	1.69	.10
5	3.71	1.44	3.57	1.55	.09
6	3.70	1.56	3.52	1.59	.11

Table 1 showed that simple word counts do not adequately replicate the scores provided by human raters. The percent of perfect + adjacent agreement ranged from .70 to .77 over the six prompts using the COUNTS model. The Pearson correlation calculated using COUNTS differed from the Pearson correlation calculated using Original scores by as much as .44. The COUNTS average Pearson correlation over the six prompts was .458 compared to the Original average of .830. COUNTS had an average discrepancy of .268 across the six prompts. This would result in nearly 27% of the scores needing adjudication.

Table 4 shows that the mean scores provided by COUNTS were much higher than the Original scores for each of the six essays with effect sizes ranging from .33 to .67.

Table 4: COUNTS Mean Scores Compared to Original Reader Mean Scores

	COUNTS		Original Reader		
Prompt	Mean	S.D.	Mean	S.D.	Effect Size
1	4.32	1.04	3.53	1.58	.60
2	4.05	1.22	3.53	1.60	.37
3	4.31	.77	3.54	1.62	.64
4	4.45	.87	3.57	1.69	.69
5	4.23	1.10	3.57	1.55	.50
6	4.19	.99	3.52	1.59	.52

Table 1 also shows that IM and COUNTS are an improvement over weighted random draws. The percent of perfect + adjacent agreement ranged from .54 to .57 over the six prompts using the PROB model. The Pearson correlation calculated using PROB differed from the Pearson correlation calculated using the Original reader scores by as much as .74. The PROB average Pearson correlation over the six prompts was .142 compared to the Original reader average of .830. PROB had an average discrepancy of .449 across the six prompts, which would result in nearly 45% of the scores needing adjudication.

Table 5 shows that the mean scores provided by PROB were higher than the Original reader scores for each of the six essays with effect sizes ranging from .18 to .31. PROB modeled the distribution of scores across all essay responses. However, the sample used in this study did not follow that distribution.

Table 5: PROB Mean Scores Compared to Original Reader Mean Scores

	PROB		Original Reader		
Prompt	Mean	S.D.	Mean	S.D.	Effect Size
1	3.82	1.03	3.53	1.58	.22
2	3.82	1.03	3.53	1.60	.22
3	3.82	1.03	3.54	1.62	.21
4	3.82	1.03	3.57	1.69	.18
5	3.82	1.03	3.57	1.55	.26
6	3.82	1.03	3.52	1.59	.31

Handling of Problematic Essays

Of the five types of "fabricated" responses, IM was able to consistently identify those labeled as copies (prompt given as the essay, repeated paragraphs, and half prompt half genuine), but the off-the-shelf model had difficulty identifying off-type and off-prompt essays.

The IntelliMetric system warning flags were numerous and specific, including flags for such things as "nonformal" words, violent language, "gobbledygook", and plagiarism. The Vantage summary of flagged items included, for each prompt, a listing of all flags by response ID and detailed listings of common text found among essays flagged with "copied prompt" and "plagiarism" flags.

With regard to the 78 fabricated responses deliberately planted into the calibration sets, the IntelliMetric system correctly identified every instance of fabricated essays involving copying, i.e., those in the "copied prompt," "repeated paragraphs," and "repeated prompt half genuine" categories. It did not fare as well on off-topic responses, but in defense of the IntelliMetric system, they were not instructed to flag off-topic essays and the issue was not part of the models they built for the evaluation.

Findings from Investigation 1

Several conclusions can be clearly drawn.

1. The Vantage IntelliMetric system automated scoring system replicates the scores provided by human raters and produces superior perfect and adjacent agreement statistics for GMAT essays.
2. The IntelliMetric system is able to identify "copied" essays.
3. The IntelliMetric system is far superior to simple word counts or simple probability modeling.
4. Very few essays would need to be adjudicated if the IntelliMetric system were to be used to verify human ratings.

In this examination, the issue of off-topic responses was not fully evaluated. Because GMAC will use the IntelliMetric system as a check against a human rater and not as a primary scoring system, the issue of off-topic responses is not viewed as a serious problem. Off-topic responses will be flagged by the human reader. An issue that was uncovered in this evaluation is that the scores provided by the IntelliMetric system were slightly higher than those provided by human raters. The differences are quite tolerable (the effect size was .15 or less). Nevertheless, Vantage worked to address the issue in the release of the IntelliMetric system 9.3, and this is an area that Vantage will be investigating and that GMAC will be watching as the IntelliMetric system goes on-line.

Investigation 2: An Evaluation Using 101 AWA Prompts

Method

A second evaluation was conducted when the initial set of 101 operational AWA prompts was calibrated for field use. As with the previous evaluation, a training set was provided to create a unique scoring model for each prompt, this time using 400 essays per prompt. Upon completion of the training and creation of an IntelliMetric system scoring model for each prompt, the scoring model was then validated using a different set of 100 essays per prompt.

The evaluation included the following calculations: Comparison of Means, Agreement Analysis, Pearson R Correlation, and a Kern Index. For the Kern Index computation, a score of +1, 0, or –2 was assigned for each essay. A score of 1 was assigned to any response where the IntelliMetric system agreed exactly with the human scores, a score of 0 was assigned to any responses where the IntelliMetric system agreed within one point, and a score of –2 was assigned if the scores differed by two or more points. The Kern Index was computed as a sum of the (Exacts*1 – Discrepants*2)/N.

Note that the index is biased towards achieving exact agreement over adjacent agreement, and it assumes that discrepancies are more egregious than exact agreements are beneficial. Using this calculation, values above .40 are generally considered acceptable, and values of .50 and above are considered more desirable.

Results

Tables 6 and 7 show the results for Analysis of an Argument prompts and for Analysis of an Issue prompts. Comparisons of means using correlated t-tests found no significant differences at $\forall = .05$ between the average human score and the IntelliMetric system score for the validation set.

Table 6: Argument Prompts: Comparison of Means

Argument Prompt ID	Percent Exact	Percent Adajecent	Perfect + Adjacent	Pearson	Kern Index	IM Mean	IM S.D.	Human Mean	Human S.D.
00001M	49%	45%	94%	.79	.37	3.92	1.11	3.85	1.34
00002M	54%	38%	92%	.67	.38	3.84	.85	4.03	1.14
00003M	51%	47%	98%	.81	.47	3.97	1.08	3.96	1.25
00005M	58%	37%	95%	.82	.48	3.63	1.11	3.72	1.29
00008M	52%	45%	97%	.73	.46	4.06	.98	4.12	1.13
00009M	43%	50%	93%	.74	.29	3.87	1.17	3.91	1.25
00021M	57%	40%	97%	.77	.51	3.88	.97	3.98	1.09
00022M	65%	32%	97%	.86	.59	3.90	1.08	3.88	1.30
00023M	60%	37%	97%	.84	.54	3.79	1.13	3.92	1.28
00024M	59%	40%	99%	.84	.57	3.90	1.07	3.88	1.22
00026M	50%	50%	100%	.82	.50	3.82	.97	3.86	1.22
00027M	42%	54%	96%	.73	.34	3.77	1.03	3.77	1.19
00029M	54%	45%	99%	.81	.52	3.91	1.03	3.90	1.19
00030M	57%	39%	96%	.81	.49	3.89	1.09	3.98	1.23
00031M	52%	43%	95%	.81	.42	3.71	1.15	3.64	1.35
00032M	59%	40%	99%	.84	.57	3.86	1.06	3.86	1.21
00033M	49%	46%	95%	.77	.39	3.77	1.06	3.89	1.24
00034M	53%	44%	97%	.79	.47	4.10	1.08	4.08	1.21
00035M	59%	39%	98%	.78	.55	3.98	1.03	4.10	1.09
00036M	44%	51%	95%	.74	.34	4.12	1.01	4.09	1.24
00037M	48%	49%	97%	.78	.42	3.81	1.04	3.90	1.25
00038M	55%	44%	99%	.85	.53	3.92	1.09	4.12	1.25
00039M	61%	36%	97%	.83	.55	3.95	1.06	4.03	1.22
00040M	54%	44%	98%	.83	.50	3.84	1.18	3.88	1.29
00041M	42%	51%	93%	.69	.28	3.95	.96	3.81	1.25
00042M	50%	44%	94%	.77	.38	3.84	1.07	3.87	1.35
00043M	65%	33%	98%	.89	.61	3.69	1.29	3.78	1.38
00044M	62%	36%	98%	.87	.58	3.81	1.16	3.83	1.35
00045M	57%	39%	96%	.84	.49	3.72	1.17	3.85	1.32
00074M	64%	34%	98%	.88	.60	3.73	1.22	3.85	1.34
00075M	60%	35%	95%	.83	.50	3.87	1.29	3.87	1.33

Table 6: Argument Prompts: Comparison of Means (continued)

Argument Prompt ID	Percent Exact	Percent Adajecent	Perfect + Adjacent	Pearson	Kern Index	IM Mean	IM S.D.	Human Mean	Human S.D.
00076M	59%	36%	95%	.84	.49	3.83	1.18	3.83	1.36
00080M	58%	37%	95%	.82	.48	3.82	1.24	3.82	1.34
00081M	46%	48%	94%	.78	.34	3.78	1.11	3.74	1.35
00082M	65%	31%	96%	.85	.57	3.68	1.20	3.78	1.35
00083M	58%	39%	97%	.84	.52	3.68	1.25	3.68	1.36
00118M	60%	40%	100%	.84	.60	4.07	1.03	4.17	1.14
00124M	58%	40%	98%	.86	.54	3.72	1.27	3.85	1.37
00126M	54%	44%	98%	.84	.50	3.88	1.15	3.90	1.32
00129M	47%	47%	94%	.76	.35	4.04	1.07	4.00	1.33
00130M	57%	42%	99%	.85	.55	3.91	1.05	3.95	1.28
00132M	61%	38%	99%	.89	.59	3.67	1.23	3.75	1.40
00135M	53%	43%	96%	.81	.45	3.75	1.17	3.92	1.23
00138M	65%	35%	100%	.91	.65	3.81	1.26	3.76	1.39
00139M	60%	38%	98%	.83	.56	3.88	1.03	3.90	1.21
00144M	55%	41%	96%	.76	.47	4.00	.85	3.97	1.15
00145M	61%	37%	98%	.87	.57	3.79	1.20	3.94	1.30
00146M	53%	41%	94%	.78	.41	3.87	1.16	3.78	1.36
00148M	52%	46%	98%	.77	.48	3.95	.94	3.97	1.15

Argument Prompt Summary

Exact Agreement Range:	42% to 65%	Average: 60%
Perf + Adj Agreement Range:	92% to 100%	Average: 98%
Kern Index Range:	.28 to .65	Average: .49
Pearson R Correlation Range:	.67 to .91	Average: .81

Table 7: Issue Prompts: Comparison of Means

Issue Prompt ID	Percent Exact	Percent Adajecent	Perfect + Adjacent	Pearson	Kern Index	IM Mean	IM S.D.	Human Mean	Human S.D.
00010M	55%	40%	95%	.74	.45	3.69	.96	3.81	1.13
00011M	60%	39%	99%	.81	.58	4.02	.99	4.13	1.10
00013M	58%	40%	98%	.86	.54	3.65	1.19	3.75	1.35
00015M	58%	40%	98%	.81	.54	4.21	1.08	4.13	1.16
00017M	64%	32%	96%	.83	.56	3.83	1.06	3.83	1.23
00018M	61%	39%	100%	.86	.61	3.81	1.07	3.92	1.22
00019M	50%	47%	97%	.74	.44	3.79	.90	3.92	1.13
00020M	66%	29%	95%	.81	.56	3.98	1.06	4.09	1.16
00046M	60%	39%	99%	.87	.58	3.71	1.21	3.80	1.32
00047M	54%	45%	99%	.85	.52	3.79	1.04	3.84	1.33
00048M	58%	40%	98%	.85	.54	3.86	1.12	3.86	1.33
00049M	69%	31%	100%	.91	.69	3.74	1.28	3.77	1.36
00050M	62%	36%	98%	.88	.58	3.77	1.23	3.81	1.39
00051M	64%	34%	98%	.87	.60	3.80	1.16	3.86	1.31
00052M	57%	43%	100%	.87	.57	3.83	1.16	3.82	1.31
00053M	63%	35%	98%	.85	.59	3.66	1.28	3.83	1.36
00054M	74%	25%	99%	.92	.72	3.86	1.24	3.81	1.32
00055M	62%	36%	98%	.85	.58	3.93	1.11	4.11	1.20
00056M	49%	47%	96%	.73	.41	4.11	1.03	4.18	1.12
00057M	54%	45%	99%	.81	.52	3.89	1.09	3.99	1.24
00058M	55%	44%	99%	.80	.53	4.12	1.02	4.14	1.15
00059M	65%	31%	96%	.81	.57	4.23	1.05	4.20	1.15
00060M	58%	38%	96%	.79	.50	3.97	1.06	4.11	1.13
00061M	58%	42%	100%	.83	.58	4.10	1.03	4.18	1.14
00062M	53%	44%	97%	.80	.47	3.96	1.08	4.02	1.22
00063M	53%	46%	99%	.79	.51	3.87	.96	3.99	1.12
00065M	64%	35%	99%	.83	.62	4.16	1.12	4.12	1.16
00066M	51%	47%	98%	.82	.47	4.00	1.24	4.01	1.23
00067M	48%	50%	98%	.77	.44	3.73	.98	3.87	1.16
00068M	49%	49%	98%	.48	.45	4.07	1.03	4.18	1.17
00069M	67%	33%	100%	.91	.67	3.79	1.20	3.84	1.36
00070M	69%	28%	97%	.88	.63	3.80	1.19	3.84	1.31

Table 7: Issue Prompts: Comparison of Means (continued)

Issue Prompt ID	Percent Exact	Percent Adajecent	Perfect + Adjacent	Pearson	Kern Index	IM Mean	IM S.D.	Human Mean	Human S.D.
00071M	64%	35%	99%	.89	.62	3.74	1.28	3.79	1.37
00072M	58%	39%	97%	.85	.52	3.83	1.29	3.86	1.33
00073M	65%	32%	97%	.88	.59	3.66	1.19	3.76	1.36
00077M	68%	31%	99%	.91	.66	3.72	1.27	3.81	1.38
00078M	60%	38%	98%	.86	.56	3.72	1.15	3.82	1.33
00079M	67%	33%	100%	.92	.67	3.66	1.34	3.79	1.37
00084M	80%	19%	99%	.94	.78	3.71	1.29	3.76	1.35
00085M	65%	34%	99%	.89	.63	3.78	1.21	3.80	1.32
00086M	70%	29%	99%	.91	.68	3.79	1.37	3.80	1.33
00087M	63%	37%	100%	.89	.63	3.80	1.19	3.83	1.33
00119M	54%	46%	100%	.89	.54	3.44	1.26	3.72	1.38
00120M	60%	38%	98%	.78	.56	3.78	.93	3.92	1.04
00125M	56%	42%	98%	.80	.52	4.18	1.09	4.26	1.15
00128M	54%	45%	99%	.85	.52	3.76	1.13	3.77	1.31
00131M	63%	36%	99%	.88	.61	3.74	1.29	3.78	1.32
00133M	67%	30%	97%	.88	.61	4.13	1.00	4.14	1.17
00137M	56%	43%	99%	.85	.54	3.73	1.04	3.78	1.30
00140M	54%	45%	99%	.79	.52	3.88	.96	3.99	1.12
00143M	56%	41%	97%	.76	.50	4.00	1.00	4.13	1.07
00149M	63%	35%	98%	.88	.59	3.75	1.19	3.84	1.34

Issue Prompt Summary

Exact Agreement Range:	48% to 80%	Average: 55%
Perf + Adj Agreement Range:	95% to 100%	Average: 97%
Kern Index Range:	.41 to .78	Average: .57
Pearson R Correlation Range:	.73 to .94	Average: .84

Findings from Investigation 2

Exact agreement ranged from 42% to 80% with an average agreement of 58%. Perfect + Adjacent rates ranged from 92% to 100% with an average agreement of 97%. The average Kern index across the Issue and Argument prompts was .53, with an index range of .28 to .78. Pearson R correlations of agreement between human raters and the IntelliMetric system averaged .83, with a range of .67 to .94.

These results confirmed findings from the previous study using only six prompts: The Vantage IntelliMetric system automated scoring system consistently calculates scores, closely matching those provided by human raters and producing reliable perfect and adjacent agreement statistics for GMAT essays. A slightly stronger match was reported for Issue prompt data than for Argument prompt data in relation to scores calculated by human raters, but concern regarding possible upward bias using the IntelliMetric system noted in Investigation 1 may be unfounded. Here, the mean differences between the IntelliMetric system and human raters fluctuate in both directions.

Discussion

In concept, a functioning model replicates the scores that would have been provided by all the human raters used in the calibration essay. Thus, a functioning model should be more accurate than the usual one or two human raters who typically assign scores. The issue, however, is how one defines a validated functioning model. The comparison data in this study involved only two or three human raters for each essay. One never knows if the human or computer is more accurate. Nevertheless, one should expect the automated essay scoring models and humans raters to substantially agree and one should expect high correlations between machine- and human-produced scores. That is what we consistently found with the IntelliMetric system.

The first investigation with six prompts did raise a question concerning possible systematic bias in the scores provided by the IntelliMetric system. GMAC and ACT could certainly live with the small magnitude of the bias, if there was indeed systematic bias; the effect sizes were in the .08 to .15 range. The second study, with 101 prompts, greatly reduced concern for systemic bias, however. No systematic bias was observed.

This evaluation found the IntelliMetric system to be an extremely effective automated essay scoring tool. GMAC will use the IntelliMetric system to validate scores provided by human raters.

Endnotes

1 The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the authors' institutions.

2 This articles was published with permission from the Graduate Management Admission Council®. An earlier version of this paper was presented by Lawrence M. Rudner and Catherine Welch at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada, April 12–14, 2005.

3 ACT™ is a trademark of ACT, Inc. AOL® is a registered trademark of American Online, Inc. Apple® is a registered trademark of Apple Computers, Inc. Codie® is a registered trademark of the Software and Information Industry Association e-rater® and ETS® are registered trademarks of the Educational Testing Service® (ETS®). GMAC®, GMAT®, Graduate Management Admission Council®, Graduate Management Admission Test® are registered trademarks of the Graduate Management Admission Council® (GMAC®). IntelliMetric℠ and My Access! are trademarks of Vantage Technologies Knowledge Assessment, L.L.C. Microsoft® is a registered trademark of Microsoft Corporation. Sun Microsystems is a registered trademark of Sun Microsystems, Inc.

References

ACT (2004). Evaluation of the GMAT® AWA Item Bank. Internal report to the Graduate Management Admission Council, August 31, 2004.

Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems, 15*(5), 35–37. Available: http://www.knowledgetechnologies.com/presskit/KAT_IEEEdebate.pdf

Elliot, S. & Mikulas, C. (2004). The impact of MY Access!™ use on student writing performance: A technology overview and four studies from across the nation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 12–16, 2004, San Diego, CA.

Haswell, R.H. (Jan, 2005). Machine Scoring of Essays: A Bibliography. Available: http://comppile.tamucc.edu/machinescoringbib

Iowa State Daily (1999). Fight the future of computer grading. Staff editorial, Feb 18, 1999. Available: http://Highbeam.com

McCallum, A., Rosenfeld, R., & Mitchell, T. (1998). Improving text classification by shrinkage in a hierarchy of classes. In ICML–98, 1998. Available: http://citeseer.nj.nec.com/mccallum98improving.html

Mitchell, T. (1997). *Machine Learning*. WCB/McGraw-Hill.

Myford, C.M. (2004). Book Review – Automated Essay Scoring: A Cross-Disciplinary Perspective, Mark D. Shermis and Jill C. Burstein, editors. *Journal of Applied Measurement, 5*(1), 111–114.

Rudner, L.M. (2004). Book Review – Automated Essay Scoring: A Cross-Disciplinary Perspective, Mark D. Shermis and Jill C. Burstein, editors. *Computational Linguistics*. *30*(2), June 2004

Rudner, L.M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *Journal of Technology, Learning, and Assessment, 1*(2). Available: http://www.bc.edu/research/intasc/jtla/journal/v1n2.shtml

Shermis, M. & Burstein, J. eds. 2003. Automated Essay Scoring: A Cross-Disciplinary Perspective. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Talento-Miller, E. & Rudner, L.M. (2005). *VSS Summary Report for 1997–2004* (RR–05–06). McLean, Virginia: Graduate Management Admission Council®.

Valenti, S., Neri F., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* Volume 2, 2003. Available: http://jite.org/documents/Vol2/v2p319-330-30.pdf,

Vantage Learning (2005, March). IntelliMetric℠ Modeling of 101 Operational Graduate Management Admission Test Prompts Summary Report. Newtown, PA.

Author Biographies

Lawrence M. Rudner is the Vice President for Research and Development at the Graduate Management Admission Council.

Veronica Garcia is the Research Writer/Editor in the Research and Development Department of the Graduate Management Admission Council.

Catherine Welch is an Assistant Vice President of Assessment Innovations at ACT, Inc.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College