J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 10, Number 3 · August 2010

Performance of a Generic Approach in Automated Essay Scoring

Yigal Attali, Brent Bridgeman, & Catherine Trapani

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 10, Number 3

Performance of a Generic Approach in Automated Essay Scoring

Yigal Attali, Brent Bridgeman, Catherine Trapani

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

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

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

Preferred citation:

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a Generic Approach in Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 10*(3). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

A generic approach in automated essay scoring produces scores that have the same meaning across all prompts, existing or new, of a writing assessment. This is accomplished by using a single set of linguistic indicators (or features), a consistent way of combining and weighting these features into essay scores, and a focus on features that are not based on prompt-specific information or vocabulary. This approach has both logistical and validity-related advantages. This paper evaluates the performance of generic scores in the context of the e-rater® automated essay scoring system. Generic scores were compared with prompt-specific scores and scores that included prompt-specific vocabulary features. These comparisons were performed with large samples of essays written to three writing assessments: The GRE General Test argument and issue tasks and the TOEFL independent task. Criteria for evaluation included level of agreement with human scores, discrepancy from human scores across prompts, and correlations with other available scores. Results showed small differences between generic and prompt-specific scores and adequate performance of both types of scores compared to human performance.

J·T·L·A

Performance of a Generic Approach in Automated Essay Scoring

Yigal Attali
Brent Bridgeman
Catherine Trapani
Educational Testing Service

Introduction

As measures of writing skill, essay writing assessments are often favored over measures that assess students' knowledge of writing conventions (for example, through multiple-choice tests), because they require students to produce a sample of writing and as such are more "direct." However, a drawback of essay writing assessments is that their evaluation requires a significant and time-consuming effort. These difficulties have led to a growing interest in the application of automated natural language processing techniques for the development of automated essay scoring (AES) as an alternative to human scoring of essays.

As early as 1966, Page developed an AES system and showed that an automated "rater" is virtually indistinguishable from human raters (Page, 1966). In recent years more systems were developed; the most prominent systems are the Intelligent Essay Assessor™ (IEA) by Knowledge Analysis Technologies™ (Landauer, Laham, & Foltz, 2003), Intellimetric (Elliot, 2001), a new version of the Project Essay Grade (PEG, Page, 1994), e-rater (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998), and e-rater V2 (Attali & Burstein, 2006). All these systems are based on computing multiple linguistic features that are designed to measure elements of writing quality, and combining these features to produce an essay score. Consequently, there are both similarities and dissimilarities in the types of features and techniques for combining these features used by these systems.

Project Essay Grade (Page, 1994) uses a regression-based approach in which a large number of linguistic features are used to predict the human scores of the essays. The first version of e-rater (Burstein et al., 1998) also used a regression approach with a large number of linguistic features, but relied on stepwise regression to select the most predictive features in each

application. E-rater version 2 (Attali & Burstein, 2006) aggregates sets of micro-features into a small and fixed set of features that cover different aspects of writing quality, mostly related to the form and structure of the essay. The features are then weighted to produce the final score, and the weights can be based a regression analysis or expert judgments of the importance of features. These aggregate features are used in a regression Similarly, the IntelliMetric system (Rudner, Garcia, & Welch, 2006; Elliot, & Mikulas, 2004; Elliot, 2003), also bases the scoring on a large number of features that are aggregated into one of several main classes. These are then aggregated using multiple unspecified statistical methods to produce the final score. Finally, the IEA engine uses Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998), a dimensionality-reduction method to represent the content of the essay as a vector in multi-dimensional space. The content score is based on the proximity of the vector to vectors of sets of pre-scored essays in the same dimensionality space. This content score, in turn, is combined with other linguistic measures such as style and mechanics features to produce the final essay score.

In the automated essay scoring literature, systems are trained and calibrated separately for each prompt (e.g., Landauer, Laham, & Foltz, 2003; Page, 1994; Rudner, Garcia, & Welch, 2006). This means that the features used, their weights, and scoring standards, may be different across prompts of the same assessment. Consequently, scores will have different meanings across prompts. With e-rater, the small and standardized feature set and the emphasis of features on form rather than content, allows for the possibility of applying the same scoring standards across all prompts of an assessment. For example, the effect of a particular grammar score on the essay score would be the same across prompts. Such a "generic" scoring approach produces standardized scores across prompts, and is more consistent with the human rubric that is usually the same for all assessment prompts, and thus contributes to the validity of scores. It also offers substantive logistical advantages for large-scale assessments because it allows scoring essays from new prompts without first training the system on each specific prompt. Attali and Burstein (2006) successfully applied a single scoring model across several prompts (see also Attali, 2007) for different writing assessments and grade levels. Moreover, Attali and Powers (2008, 2009) extended the notion of the generic model across assessments and ability levels, by creating a developmental writing scale based on e-rater features, a single scoring model (and standards) for timed writing performance of children from 4^{th} to 12^{th} grade.

The purpose of this paper was to compare performance results of human scores with generic and prompt-specific e-rater scores. Three types of e-rater scores were evaluated. First, generic scores were based on the

non-content features (see method section below; the content features are prompt-specific by nature) and were computed on essays written to new prompts that were not encountered in training the generic scores. Second, prompt-specific scores were based on the same non-content features, but were computed on essays written to the same prompt they were trained on. Lastly, a second type of prompt-specific scores were based on both the non-content and content features. The last two types of scores were used to separate the effects of prompt-specific training (versus generic training) and of content features (over non-content features) on performance.

The writing assessments that were analyzed are the argument and issue tasks of the analytic writing section of the Graduate Record Examination General Test (GRE) and the independent task of the Test of English as a Foreign Language internet-based test (TOEFL iBT). These assessments provide interesting contrasts in the context of this paper. The GRE issue and TOEFL independent tasks ask examinees to present their own opinions on a given topic, but the TOEFL assessment is taken by examinees whose first language is not English, and therefore may present special difficulties for automated scoring. The GRE argument task ask examinees to present a critique of a given argument, therefore its (human) scoring is more dependent on the specific content and arguments presented in the essay, which may present greater difficulties to an automated scoring system that is based on non-content features.

The question that guided this study was how the performance of the generic scores compared with performance of the prompt-specific scores. Specifically, are there enough differences in human scoring standards and student essays across prompts, that would result in lower performance of a generic approach that ignores such differences, compared with a prompt-specific approach that takes these differences into account. This question was addressed with two different prompt-specific scores, a range of writing tasks, and different populations of writers.

Method

E-rater features

The feature set used with e-rater (see Attali & Burstein, 2006, for a thorough description) includes measures of grammar (e.g., subject-verb agreement errors), mechanics (e.g., spelling errors), usage (e.g., article errors and homophone errors), style (e.g., overused words and very long or very short sentences), organization (based on the number of discourse elements, such as introduction, main ideas, and supporting ideas), development (based on the number of words per discourse element), vocabulary

(based on the level of vocabulary words used), and average word length. All these measures are related to the form and structure of the essay. Two additional content features are based on measuring the similarity of the essay vocabulary to the vocabulary of essays (from the same prompt) at different points on the score scale are sometimes used.

Data

For the GRE assessment, a random sample of up to 3,000 essays (if available) for each of the 113 issue prompts and 139 argument prompts was drawn from the available test records from September 2006 to September 2007. The minimum number of essays for a prompt was 750, and the median was 2,985. For each essay, several variables were available for analysis, among them human rater scores (at least two) on a 6-point scale, all GRE test scores of the test taker who wrote the essay, and the test taker's answers to the biographical questionnaire.

For the TOEFL assessment, the analyses include all essays written worldwide from the beginning of October 2006 until the middle of May 2007. The total number of test records was 205,566. In this period, 26 independent prompts were administered, ranging in the number of test takers from around 3,900 to around 15,000. For each test taker, several variables were available for analysis, among them the human essay scores (on a 5-point scale), all TOEFL test scores, and the test takers' answers to the biographical questionnaire.

E-rater Scoring

All automated scores were based on a regression analysis for the prediction of the human scores from the e-rater features. For generic scoring and the first type of prompt-specific scoring, only the eight non-content features were used. For the second type of prompt-specific scoring, the prompt-specific vocabulary usage features were added to the non-content features. The final e-rater scores were scaled such that their standard deviation was equal to that of a single human score. Therefore, all scores had the same standard deviation.

In order to evaluate the e-rater scores, data from each prompt was partitioned into a training and validation set (500 essays for GRE argument and issue, and 50% of the essays for TOEFL, with content features limited to 500 essays). Generic scores were produced using a prompt-fold approach. For each prompt, a single training set was created by combining all training sets excluding the one for the particular prompt (for TOEFL a subset of these essays were used in order to have an equal number of essays from each prompt). A single regression analysis was conducted on the combined training set, and the regression parameters were then applied

to score the validation set of the particular prompt. For prompt-specific scoring, a separate regression analysis was performed for each prompt on the training sample of the prompt, and the regression parameters were then applied to score the validation set of the prompt. All results below are based on scores from the validation sets.

Results

Table 1 presents the average standardized weights for the prompt-specific with content features (PSWC) scores, expressed as percentages of the sum of all weights. The table shows a similar pattern of weights across the three assessments, with a weight of around 30% for organization and development, and a weight of around 5% for each of the other six non-content features. The largest difference between assessments is the higher weight of the content features for the GRE argument task. The weights for the generic (G) scores and prompt-specific without content features (PSNC) were very similar to the average PSWC weights, with the weights of the content features dispersed between all the non-content features.

Table 1: Average Relative Weights for Prompt-Specific with Content (PSWC) Scores

	GRE Argument	GRE Issue	TOEFL Independent
Organization	34%	30%	31%
Development	26%	28%	28%
Grammar	4%	4%	7%
Usage	5%	9%	7%
Mechanics	4%	8%	9%
Style	1%	1%	2%
Vocabulary	3%	4%	3%
Word Length	3%	5%	6%
Content	20%	11%	6%

Table 2 presents agreement results between the first human score of each essay (H1) and between the second human score (H2), G, PSNC, and PSWC e-rater scores. Agreement indices include quadratic-weighted Kappas (e-rater scores were rounded to compute this statistic), product-moment correlations, and the absolute value of standardized discrepan-

cies (or effect size, *d*) between the two scores (*d* standardized with the *SD* of H1). The use of the absolute value of *d* prevents positive discrepancies on some prompts canceling negative discrepancies on other prompts. The quadratic-weighted Kappas and correlations are closely related, but correlations do not take biases between scores into account.

Table 2: Agreement of H2 and E-rater Scores with H1 (and SD across Prompts)

	H2	G	PSNC	PSWC
GRE argument (N=139)				
Weighted Kappa	.78 (.02)	.72 (.02)	.73 (.02)	.76 (.02)
Correlation	.79 (.02)	.76 (.02)	.76 (.02)	.79 (.02)
d (absolute value)	.02 (.01)	.10 (.07)	.03 (.02)	.02 (.02)
GRE issue (N=113)				
Weighted Kappa	.74 (.02)	.76 (.02)	.76 (.02)	.77 (.01)
Correlation	.74 (.02)	.79 (.01)	.79 (.01)	.80 (.01)
d (absolute value)	.02 (.02)	.05 (.04)	.03 (.02)	.03 (.02)
TOEFL independent (N=26)				
Weighted Kappa	.70 (.03)	.72 (.03)	.73 (.02)	.73 (.02)
Correlation	.70 (.03)	.76 (.02)	.76 (.02)	.77 (.02)
d (absolute value)	.02 (.02)	.07 (.05)	.01 (.01)	.01 (.01)

The table shows similar results for GRE issue and TOEFL independent tasks. In terms of Kappas and correlations, higher agreement was found between H1 and e-rater scores than between H1 and H2 (by around .06 for correlations), no difference between G and PSNC scores, and a minimal advantage of PSWC over PSNC and G scores (by around .01 for correlations). For GRE argument, human-human agreement compares relatively better with human-machine agreement, although H2 and PSWC correlations are similar. As with the other assessments, the difference between G and PSNC scores is minimal. However, in this assessment the difference between G/PSNC and PSWC scores is larger, around .03.

In terms of mean differences across prompts, both PSNC and PSWC scores show minimal discrepancies from human scores in all three assessments, with an average discrepancy of around .02 standard deviations. This is to be expected, as prompt-specific scores are trained to predict human

scores on a prompt-basis. Generic scores, on the other hand, could and do indeed show discrepancies from human scores across different prompts, with somewhat larger average d values for GRE argument (.10) than for the other two assessments (.05 and .07).

The discrepancies between generic and human scores across prompts can be evaluated by comparing these essay scores to other subscores of the same assessment, the GRE verbal scores and the reading, speaking, and listening TOEFL scores. These scores can serve as an anchor test because they are equated across administrations, whereas the essay scores across administrations (prompts) are not linked in any statistical way. In particular, if all scores (the anchor scores, human essay scores, and generic scores) are standardized across all observations in the dataset, the essay scores can be used to predict the anchor scores separately for each prompt, and the regression intercept would then signify the predicted anchor score for an average essay score. A positive (negative) intercept signifies lower (higher) scoring standards relative to other prompts, because it means that a higher (lower) essay score predicts the same average verbal score. In the context of the score linking framework developed by Holland and Dorans (2006), this kind of prediction of anchor scores from essay scores is the most basic test linking method (the other two being scale aligning and equating).

The analysis was performed on the GRE argument essays by standardizing the human (average of H1 and H2) and generic scores, and then regressing each of the standardized scores on the standardized GRE verbal scores, separately for each prompt. The difference between the absolute values of the intercepts (generic minus human) is a measure of the advantage of the human intercepts: a positive value signifies that the human intercept was closer to 0 than the G intercept. Results showed no difference between these absolute values (M=.005, SD=.054, $t(138)$=1.04, p=.30). These results show that the larger discrepancies between human and generic scores across prompts are not necessarily reflected in larger discrepancies with equated GRE verbal scores. Similar results were obtained for GRE issue and for TOEFL independent (using the average of the reading, listening, and speaking scores as an anchor).

Table 3 presents correlations of the e-rater scores and the H2 score with other scores from the same assessment, and for GRE, the self-reported undergraduate major GPA. In all cases, the e-rater correlations are slightly higher than those of H2. In most cases, G and PSNC correlations are the same and PSWC correlations are slightly higher (by .01) than PSNC correlations.

Table 3: Correlations of Essay Scores with Other Scores

	H2	G	PSNC	PSWC
GRE argument				
GRE-Verbal	.55	.56	.57	.59
GRE-Quantitative	.22	.27	.26	.26
Issue essay score	.62	.69	.69	.70
Undergraduate Major GPA	.19	.21	.21	.21
GRE issue				
GRE-Verbal	.51	.53	.54	.55
GRE- Quantitative	.07	.14	.14	.11
Argument essay score	.60	.65	.65	.66
Undergraduate Major GPA	.15	.18	.18	.18
TOEFL independent				
Reading	.56	.60	.61	.62
Listening	.58	.59	.59	.60
Speaking	.61	.61	.62	.63
Integrated score	.59	.61	.62	.63

Discussion

There are two main aspects to the generic approach for automated essay scoring. First, generic scores are standardized across writing prompts. In other words, across different prompts scores are based on the same information (features) and the same standards are used for interpreting this information. This means that essay scores can be readily compared across prompts. Second, in the generic approach the information used to evaluate essay quality is related to how the essay was written and not to what was written. That is, in the generic approach the specific content of the essay is not taken into account in its evaluation.

The first aspect of the generic approach has obvious advantages for a large-scale assessment, because it enables the comparison of essay scores across forms, and is easier to maintain when many test forms are in use. Nevertheless, this approach was not adopted in the past by AES systems.

This could be the result of an emphasis on optimal prediction of human scores, as any prediction system that takes into account prompt identity must be at least as successful as a system that does not consider prompt identity. However, it could also be the result of an emphasis on features that take into account the prompt-specific vocabulary, or content of the essay.

However, this study could not find support for the importance of neither prompt idiosyncrasies nor prompt-specific vocabulary. First, the mean differences between generic (G) and human scores across prompts were very small in most cases and across the three writing assessments. In addition, there were almost no differences in performance between G and PSNC scores, which differ only in that PSNC scores are optimized at the prompt level. These results indicate that the *human* scoring standards across prompts were consistent and therefore there is little or no advantage in predicting human scores at the prompt level.

Second, across the three different college-level writing assessments that were analyzed in this study, content, or prompt-specific vocabulary usage had only a minor effect on performance. Differences in performance between PSNC and PSWC scores, which differ only in the PSWC scores' use content features, were noted only for GRE argument. These results indicate that content does not have an important role in *human* scoring of these types of writing assessments.

This finding may be counter-intuitive to many raters. For example, Ben-simon and Bennett (2007) asked raters to assign importance weights to different dimensions of essay quality and found that the dimension of *topical analysis* received a high weight (around 30% on average). Huot (1988) coded talk-aloud protocols of raters and found that most comments were related to content and organization. Similarly, Breland and Jones (1984) had essays scored holistically by raters and also annotated for their strong and weak points and concluded that discourse characteristics had more influence on rater judgments than syntactic and lexical characteristics. However, as the results of this study show, the perceptions of raters do not necessarily correspond to the relations between human ratings and objective measures of essay quality.

Yet, the importance of content in essay ratings is not fundamental to the generic approach. It is possible to conceive of generic scores that integrate content features, albeit with a single scoring standard across prompts, and there are surely the human ratings of other tasks or assessments could be more dependent on content. On the other hand, standardization of scoring standards provides important advantages for AES, above all extending generic scores beyond a single writing assessment. The

key to the developmental writing scale (Attali & Powers, 2008, 2009) was to use objective scoring and the same scoring standards across prompts and grade levels, in order to provide a standardized measurement across the developmental range.

Similarly, note that the GRE issue and TOEFL independent assessments, who share similar tasks and prompts, have similar feature weights (Table 1, page 8), despite the enormous differences in English language capacity of their respective examinee populations. Moreover, many of the TOEFL examinees also take the GRE. However, whereas with human scoring (or prompt-specific automated scoring) comparing scores across two assessments requires a special scaling study, generic scoring easily provides this capability. The simplicity with which comparability of scores can be achieved with generic scoring can be applied in other situations. For example, generic scores can help large-scale assessments manage scoring standards across time and raters. Generic scoring can also help teachers interpret their own scoring standards in comparison with state or other assessments.

In summary, the generic approach has important implications for the validity and acceptability of AES as it produces more interpretable and thus meaningful scores. However, it also provides new possibilities for application beyond predicting the scores of human raters.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org.

Attali, Y. (2007). *Construct Validity of e-rater in Scoring TOEFL Essays* (ETS RR-07-21). Educational Testing Service: Princeton, NJ.

Attali, Y., & Powers, D. (2008). *A developmental writing scale* (ETS RR-08-19). Educational Testing Service: Princeton, NJ.

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and Psychological Measurement*, 69, 978–993.

Ben-Simon, A., & Bennett, R.E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment, 6*(1). Available from http://www.jtla.org.

Breland, H. M., & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication*, 1, 101–119.

Written Communication, 1, 101-119. Burstein, J.C., Kukich, K.,Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Elliot, S.M. (2001, April). *IntelliMetric™: From here to validity*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Elliott, S. (2003). *Intellimetric™: From here to validity*. In M.D. Shermis & J.C. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 71–86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elliot, S., & Mikulas, C. (2004, April). *How does IntelliMetric™ score essay responses? A mind based approach*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Holland, P., & Dorans, N. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Huot, B. (1988). *The validity of holistic scoring: A comparison of the talk-aloud protocols of novice and expert holistic raters*. Unpublished PhD dissertation, Indianna University of Pennsylvania.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: *A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Page, E.B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 48, 238–243.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62, 127–142.

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

Author Biographies

Yigal Attali is a Senior Research Scientist at Educational Testing Service. He is interested in exploring cognitive aspects of assessment and in the implementation of technology in assessment. His current research includes development and evaluation of automated essay scoring and assessment of feedback mechanisms in constructed response tasks. He received his B.A. in computer sciences and his Ph.D. in psychology from the Hebrew University of Jerusalem. Yigal Attali can be reached at yattali@ets.org.

Brent Bridgeman received his Ph.D. in educational psychology from the University of Wisconsin, Madison. He joined Educational Testing Service in 1974 after several years of college teaching; his current title is Distinguished Presidential Appointee. In addition to over forty publications in refereed journals and seven book chapters, he has authored dozens of ETS Research Reports and made numerous presentations at national and international conferences. Dr. Bridgeman's recent work focuses on validity and fairness issues related to test question formats, test time limits, and automated scoring of spoken and written responses.

Catherine Trapani is a Principal Research Data Analyst at Educational Testing Service (ETS) in Princeton, NJ. Cathy has worked on a variety of projects within Research and is currently responsible for the operational integrity of the automated scoring engines for such clients as GRE(r) and TOEFL(r). Other research interests include quality of human scoring, teacher effectiveness, using exploratory data analysis and graphical methods to present data in simpler displays, and applying decision theory to judgments. She earned an M.S. in statistics from Montclair State University and a B.A. in Physics from Rutgers College. She is currently studying for her Ph.D. in Psychometrics at Fordham University.

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