GRE®
RESEARCH

Comparing the Validity of Automated and Human Essay Scoring

Donald E. Powers
Jill C. Burstein
Martin Chodorow
Mary E. Fowles
Karen Kukich

June 2000

GRE Board Research Report No. 98-08aR

ETS Research Report 00-10

ETS

Educational Testing Service, Princeton, New Jersey 08541

Comparing the Validity of Automated and Human Essay Scoring

Donald E. Powers
Jill C. Burstein
Martin Chodorow
Mary E. Fowles
Karen Kukich

GRE No. 98-08aR

June 2000

This report presents the findings of a
research project funded by and carried
out under the auspices of the
Graduate Record Examinations Board

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

Acknowledgments

Special thanks go to Chi Lu for generating *e-rater* scores, to Inge Novatkoski for analyzing our data, to Ruth Yoder for help in preparing this report, including the graphics it contains, and to Leona Aiken, Rich Swartz, and Art Young for providing helpful reviews of an earlier draft.

Abstract

This study sought to provide further evidence of the validity of automated, or computer-based, scores of complex performance assessments, such as direct tests of writing skill that require examinees to construct responses rather than select them from a set of multiple choices. While several studies have examined agreement between human scores and automated scoring systems, only a few have provided evidence of the relationship of automated scores to other, independent indicators of writing skill. This study examined relationships of each of two sets of Graduate Record Examinations (GRE®) Writing Assessment scores -- those given by human raters and those generated by *e-rater* (the system being researched for possible application in a variety of assessments that require natural language responses) -- to several independent, nontest indicators of writing skill, such as academic, outside, and perceived success with writing.

Analyses revealed significant, but modest, correlations between the nontest indicators and each of the two methods of scoring. That automated and human scores exhibited reasonably similar relations with the nontest indicators was taken as evidence that the two methods of scoring reflect similar aspects of writing proficiency. These relations were, however, somewhat weaker for automated scores than for scores awarded by humans.

Keywords: Writing assessment, writing skills, Graduate Record Examinations (GRE®), validity, automated scoring, essay scoring

Table of Contents

List of Tables

List of Figures

Introduction

As measures of writing skill, so-called "direct" assessments (i.e., those that require examinees to *produce* a sample of writing) are usually favored over indirect measures (i.e., those that survey writers' *knowledge* of particular writing conventions). This preference results, presumably, from the fact that the writing skills of most interest -- the ability to sustain a well-focused, coherent discussion, for example -- are displayed only in direct assessments, not in their less direct counterparts. Despite this capability, direct measures are not always used because they suffer from at least one unmistakable limitation: Whereas indirect measures yield answers (e.g., multiple-choice selections) that are easily scored by machine, the responses to direct assessments (e.g., extended, natural-language essays) typically require more labor-intensive, and therefore expensive, evaluation efforts.

One hope for reducing the costs associated with direct assessments of writing comes from efforts by computational linguists to evaluate essays automatically by using computer-assisted methods. Although this research can be traced back at least 30 years (e.g., Daigon, 1966; Page, 1966, 1968), automated scoring has, to our knowledge, yet to be fully implemented (i.e., used to the exclusion of human readers[1]) in any major high-stakes testing program. And perhaps it never will, for it is difficult to envision how a computer could ever judge all of the qualities that constitute effective writing. At least one major writing test, however -- the Graduate Management Admission Council's (GMAC) Analytical Writing Assessment -- uses both automated and human scoring in combination. For this assessment, each examinee essay is scored both automatically and by one human reader (instead of by two human readers, as has been the tradition for most high-stakes writing assessments).

Reactions to the Automated Scoring of Essays

There is little disagreement about the computer's ability to produce evaluations of writing that are, unlike those provided by human readers, always highly *reliable.* As has been noted (Schwartz, 1998, for example), computers are not subject to any of the traits or conditions, such as impatience or fatigue, that render human evaluators less than perfectly consistent over time. Moreover, if programmed in the same way, different computers will, unlike their human counterparts, virtually always agree exactly with one another.

There is, however, considerably more contention about the *validity* of automated essay scoring (Murray, 1998; Labi, 1999; Scott, 1999). Some critics have argued that automated scoring tends to focus on the writer's grammar and use of the conventions of standard written English. Therefore, the argument goes, direct assessments based on these methods of evaluation are no more informative than assessments employing less direct multiple-choice questions (DeLoughry, 1995; Mitchell, 1998). If true, these beliefs suggest that automated scores may, in the parlance of test validation, underrepresent important aspects of the construct of interest -- writing proficiency (Messick, 1989). It is likely, therefore, that if automated scoring methods are to be accepted widely, compelling evidence will be needed to support their use, for the way in which an assessment is evaluated is an integral aspect of its construct validity (Messick, 1995). The goal of the research described here was to provide further evidence of the validity of automated scores generated by one specific computer-based method, *e-rater*, the system being researched for possible application in a variety of assessments that require natural language responses.

Relevant Research

Agreement Between Automated and Human Scores

Despite the diverse opinions about automated scoring, steady progress has been made in developing methods to analyze natural language responses. Moreover, these methods have now undergone several evaluations to assess their accuracy. Thus far, however, most of these studies have focused on agreement -- that is, the correspondence between automated scores and those assigned by trained human readers. For example, Petersen (1997) reported that Page's automated scoring method (Page & Petersen, 1995) was able to generate scores that agreed strongly, but not perfectly, with those awarded by human readers. Petersen's investigation focused on expository essays written by prospective teachers for a teacher certification program, the Praxis Series: Professional Assessments for Beginning Teachers™. The results revealed a median correlation of .65 between pairs of human readers, while the median correlation between computer scores and human scores was .72. That is, the computer exhibited greater agreement with human readers than human readers did with one another. For another set of essays that required the discussion of an issue, the same researchers also found the correlation between automated and human readers' scores (r = .75) to be as strong as the correlation between pairs of human readers. Again, however, it should be noted that none of the relationships -- neither between human readers, nor between automated and human readers -- was perfect.

Similar results have been found in other studies. For example, using an alternative (i.e., non-*e-rater)* model of automated scoring to analyze the teacher certification essays studied by Page and Peterson (1995), Kaplan, Wolff, Burstein, Lu, Rock, and Kaplan (1998) were able to replicate Page and

Peterson's results. More recent research on the Graduate Management Admission Test (GMAT®) has demonstrated comparable results for essays written by applicants to graduate schools of management (Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock, & Wolff, 1998). Using the *e-rater* model, the latter study examined each of the two different kinds of essays ("discuss an issue" and "analyze an argument") that are used on the GMAT writing assessment. Finally, again using the *e-rater* model, comparable levels of agreement have been found between automated scores and scores by human raters for essays written by nonnative speakers of English for the Test of Written English, a test designed to assess the writing skills of international students (Burstein & Chodorow, 1999). The correlation was .69, and the percent of matches that were exact or within one point of one another reached 92% (chance agreement was 44%).

Relationships With Other Criteria

In addition to studies of the agreement between human raters and automated scoring systems, a few studies have provided evidence of the relationship of automated scores to other important criteria. For instance, for both Praxis Series essays and essays based on the Graduate Record Examination (GRE®) Writing Assessment, Petersen (1997) reported correlations with several other independent variables. For Praxis essays, the correlations of automated scores with scores from other subtests in the Praxis Series (reading, mathematics, and writing based on editing and error recognition tasks) were .39, .30, and .47, respectively. The corresponding correlations for scores assigned by human readers were .43, .36, and .45. For GRE essays, the correlations with GRE verbal, quantitative, and analytical scores were .26, .13, and .24, respectively, for automated scores, and .33, .16, and .26, respectively, for human readers. That human and automated scores relate in a similar manner to other variables suggests that both may reflect similar constructs.

In another effort, Landauer, Laham, Rehder, and Schreiner (1997) used latent semantic analysis (Landauer, Foltz, Laham, 1998) to evaluate student essays. Both automated scores and human scores were related to an external criterion consisting of a short-answer test covering writers' understanding of the essay topic. The correlation with performance on the criterion test was .70 for human raters and .81 for automated scores.

From the preceding discussion it is clear that there is some evidence, albeit limited, comparing the relationships of human and automated scores to other, independent indicators of writing skill. The aim of the study reported here was to generate more such evidence.

Method

Description of the GRE Writing Assessment

The GRE Writing Assessment is designed to measure examinees' ability to:

- articulate complex ideas clearly and effectively
- examine claims and accompanying evidence
- support ideas with relevant reasons and examples
- sustain a well-focused, coherent discussion
- control the elements of standard written English

The assessment consists of two writing tasks: a 45-minute "issue" task that requires examinees to present their perspective on an issue, and a 30-minute "argument" task that requires them to analyze an argument. The prompt for the issue task states an opinion on an issue of general interest and asks test takers to address the issue from any perspective(s) they wish, providing relevant reasons and examples to explain and support their views. The prompt for the argument task presents an argument and requires test takers to critique it, discussing how well reasoned they find it, rather than simply agreeing or disagreeing with the position expressed in the argument. The two tasks are intended to complement one another: One requires test takers to construct their own arguments by making claims and providing evidence to support a position; the other requires the critique of someone else's argument. More detail about the GRE Writing Assessment and how it is scored is available at http://www.gre.org/twotasks.html as well as in Schaeffer, Fowles, & Briel (in press).

Description of *E-Rater*

The focus of our study was *e-rater,* the automated scoring system that has been developed to evaluate test takers' responses to essay questions, like those that constitute the GRE Writing Assessment. In brief, *e-rater* uses natural language processing techniques to model the performance of human evaluators. For each essay prompt, a sample of essays is selected to represent the complete range of possible scores. The system is then "trained" on these sample essays, which have been previously scored by human readers.

In its attempt to model human readers, *e-rater* first uses several subroutines to extract a wide variety of features from test takers' essays. These features are then used in combination to predict the scores that were assigned previously by human readers to the essays in *e-rater's* training sample. Currently, ordinary least squares linear regression analysis is used to select, in a stepwise fashion, the set of features that best predicts these scores.

In total, 50-60 features are "extractable" from test takers' essays, but in practice only a subset of the most predictive features, usually about 8 to 12, is retained and used for each essay prompt. The features that *e-rater* uses are required to be predictive of human readers' scores, but they must also have some correspondence to the characteristics that human readers are trained to consider when scoring essays. That is, they must be related to the qualities or features -- coherence, for example, that are emphasized in the GRE Writing Assessment scoring guide. The least squares regression weights for these predictors are applied to each new essay to estimate a score, which is rounded to the nearest integer from 0 to 6 in order to place it on the scale used by human readers.

E-rater's focus is on three general classes of essay features:

- structure -- indicated by the syntax of sentences
- organization -- indicated by various rhetorical features that occur throughout extended discourse
- content -- indicated by prompt-specific vocabulary

Structural characteristics include syntactic variety -- that is, the use of various structures in the arrangement of phrases, clauses, and sentences. This category also includes such features as the number of subjunctive modal auxiliary verbs (would, could, should, might, may), the prevalence of infinitive clauses, and the incidence of subordinate clauses.

E-rater evaluates test takers' ability to organize their ideas by extracting a variety of rhetorical features -- that is, characteristics that are associated with the orderly presentation of ideas. Of special interest here are cue words, or terms that signal where an argument begins and how it is developed. (In this case, when we say "argument," we mean any rational presentation that uses reasons or examples to persuade the reader.) For example, terms such as *in summary* and *in conclusion* are used for summarizing, words like *certainly* and *presumably* are used to express opinions or beliefs, and other words or phrases signal still other aspects of an argument. In short, *e-rater* looks for linguistic clues -- such as transitional words or logical connections between sentences and clauses -- that signify analytical or well-ordered thinking.

By means of topical analyses, *e-rater* also considers an essay's content -- that is, prompt-specific vocabulary. These analyses are predicated on the assumption that, compared with essays that are poorly written, well-written essays tend to use more precise and specialized topic-related vocabulary. The expectation is that, with respect to the words they contain, better essays will bear a greater resemblance to other good essays than to poorer ones. Weak essays, on the other hand, will be similar to other weak essays. Acting on this assumption, *e-rater* evaluates each essay and assigns a score based on the similarity of its content to samples of previously scored essays.

Besides evaluating an essay's content as a whole, *e-rater* also considers an essay's content argument-by-argument. The rationale is that additional information can be extracted about an essay's content by examining clusters of word groupings, in this case individual arguments. In a manner similar to that used for essays as a whole, a content score is assigned to each argument on the basis of how well an essay's arguments match the content of essays scored previously by human readers.

The final step is to use all of the features that *e-rater* extracts (or rather, the values that it assigns to these features) to predict the scores assigned by human readers. A model -- that is, a set of features that is most predictive of human readers' scores -- is specified for each essay prompt. The set is generally somewhat different for each prompt. More detail about how *e-rater* functions is available in a number of reports that can be downloaded from the following Web site: http://www.ets.org/research/erater.html. See also Burstein et al. (1998).

Data Source

The analyses reported here are based on data collected for two previously reported studies of the GRE Writing Assessment (Powers, Fowles, & Welsh, 1999; Schaeffer, Fowles, & Briel, in press). For these earlier studies, more than 2,000 prospective graduate students were recruited from 26 geographically diverse colleges and universities. Each of these participants composed two GRE essays at a test center. Half of the group wrote one issue and one argument essay, and the other half wrote either two issue or two argument essays. Each participant also provided other kinds of information -- referred to as nontest indicators -- about their writing skills. These nontest indicators, which served as validity criteria, included:

- two samples of writing prepared for undergraduate course assignments -- one piece ("Sample X") that was typical of their writing and another ("Sample Y") that was of somewhat lower quality[2] -- each scored on a 1 to 6 scale
- self-evaluations of writing ability (in comparison to peers)
- self-reported grades in undergraduate courses that required considerable writing
- self-reported, documentable accomplishments in writing (e.g., publishing a letter to the editor, authoring a paper for a professional meeting, etc.)
- self-reported success with various *kinds* of writing (e.g., research papers, persuasive writing, etc.) and with various *processes* of writing (e.g., organizing, drafting, and editing/revising)

Correlations among these various criteria were quite modest, suggesting that each reflects a somewhat different aspect of a more general construct of writing proficiency. Of course, each is also less than perfectly reliable, with reliability estimates (and method of estimation) as follows:

- .19 for self-evaluations of writing ability (test-retest)
- .40 for a single writing sample (.57 for two) (correlation between samples)
- .63 for reports of success with various *kinds* of writing (test-retest)
- .71 for reports of success with various *processes* of writing (test-retest)
- .84 for reports of accomplishments in writing (internal consistency)

Complete and usable data were available for approximately 1,700 of the study participants. More detail -- such as information about reliability calculations for each of the nontest indicators and how the course-related writing samples were evaluated -- is available in Powers et al. (1999).

Procedures

Our database contained 101 to 149 essays for each of 20 argument prompts and 20 issue prompts that were administered to the sample. Two trained human readers had scored each of these essays previously. A major task for the current study was to generate for each essay in the database an *e-rater* score to accompany the scores previously assigned by human readers. However, to date, it has been deemed necessary to base *e-rater*'s training on at least 250 essays for each essay prompt -- about twice as many as were available for our study. In order to deal with the lack of sufficient training data, two alternative procedures were implemented.

First, we developed general training models that were not specific to individual prompts. To implement this strategy, we pooled all issue essays, regardless of the particular prompts on which they were written, and all argument essays, again regardless of prompts. Next, we developed a single, more general model for each of these two kinds of prompts. This procedure seemed reasonable, as the prompt-specific models developed to date have tended to use many of the same features. Thus, the quality of *e-rater* scores would not degrade significantly, we hoped, by using models that weren't tailored precisely to each prompt.

Second, in order to improve the stability of our results and to guard against overfitting the data, we also employed a jackknife technique to make maximum use of the number of essays available. For each prompt, a model was developed on the basis of n-1 of the essays written for the prompt, and based on this model, an *e-rater* score was computed for the remaining essay. This procedure was repeated n times for each essay prompt in order to get an *e-rater* score for each essay. Similarly, a score was obtained for each essay written for each of the other prompts.

Results

Interreader Agreement

The correlation between individual human readers was .85 for issue essays and .83 for argument essays. Across all prompts, pairs of human readers agreed very highly with one another -- exactly 68% of the time, and either exactly or within one point of one another 99% of the time. Scores based on *e-rater* general models agreed exactly with individual human readers 48% of the time, and exactly or within one point 94% of the time. The comparable statistics for agreement between prompt-specific jackknife scores and human reader scores were virtually identical (48% and 93%). To put these agreement rates into perspective, the "generic" baseline chance agreement on a six-point scale, such as the one on which GRE essays are scored, is 17% (exact) and 44% (exact or within one point). Chance rates for each pair of data sets would be somewhat higher if we assume that readers mimic the true distribution of essay scores.

Because GRE Writing Assessment scores are based on one issue and one argument essay, the remaining results have been restricted to a sample of approximately 900 study participants who wrote one essay for each of these two kinds of prompts. For this sample, the correlation between scores assigned by human readers (summed over four readers in all, two for each essay) and *e-rater* jackknife scores (summed over two jackknife scores, one for each essay) was .74. The correlation between human reader scores and *e-rater* general model scores (summed over two general model scores, one for each essay) was

.73. The correlation between *e-rater* scores based on general models and those based on the jackknife method was .82.

Correlations with Nontest Indicators

E-rater scores based on the jackknife method and the general models bore very similar patterns of correlations with the nontest indicators -- as would be expected on the basis of the strong relation between scores from the two methods. Because a somewhat greater number of scores was available for the jackknife method, we present data here only for that method. (Results based on the general models are available from the authors, as are analyses based on the smaller numbers of participants who wrote either two issue essays or two argument essays, which yielded generally similar, although slightly lower, patterns of correlations.)

Table 1 displays correlations of GRE Writing Assessment scores (i.e., the average of scores on the issue and argument essays) with participants' status on each of the nontest indicators of writing for which information was collected. Correlations are given for each of several combinations of human and automated scores. As found earlier (Powers et al., 1999), performance on the GRE Writing Assessment (as evaluated by trained human readers) correlates modestly with each of several nontest indicators of writing proficiency.[3] Correlation coefficients are highest for a criterion based on samples of students' undergraduate writing, and weakest for a criterion based on students' reports of their writing-related accomplishments. With one exception, *e-rater* scores correlated slightly less strongly with each of the nontest criteria than did the scores assigned by human readers. Using a test for the difference between correlated correlations (Meng, Rosenthal, & Rubin, 1992), we found the differences to be statistically significant at the .05 level or beyond for at least half of the indicators. The rank order of correlations with the nontest indicators was similar for *e-rater* and human scoring, and the median correlation over all indicators was nearly the same for *e-rater* ($\underline{r} = .16$) and human scores ($\underline{r} = .18$). That is, the indicators that relate most (or least) strongly to human scores also relate most (or least) strongly to *e-rater* scores. Finally, and potentially of most significance, we note that when *e-rater* scores are combined with those from human readers (regardless of whether one or two human readers are employed), there is relatively little decrement in validity, relative to estimates based solely on two human readers.

There are several possible reasons why *e-rater*'s performance is below that of its human counterparts. For instance, the lower correlations may be partly a function of *e-rater* scores being somewhat less variable than the scores assigned by human readers. While the means for human scores and *e-rater* scores are approximately the same (3.8 vs. 3.7 for issue essays, and 3.3 vs. 3.3 for argument

Table 1

Correlations of GRE Writing Assessment Scores with Nontest Indicators of Writing Skill, for Human and Automated Scoring

Indicator	One human reader#	Two human readers	*E-rater*	One human reader and *e-rater*	Two human readers and *e-rater*
Writing samples (readers' grades)[1]	.31	.38	.24	.33	.36
GPA in writing courses[2]	.29	.34	.27	.33	.34
Self comparison with peers[3]	.24	.29	.17	.25	.27
Success with various kinds of writing[4]	.22	.26	.16	.23	.25
GPA overall	.18	.20	.17	.20	.21
Success with writing processes[5]	.17	.20	.13	.18	.19
GPA in major field	.12	.14	.13	.15	.15
Writing samples (professors' grades)[6]	.13	.16	.12	.15	.16
Accomplishments (log)[7]	.06	.07	.09	.08	.08

[1] Two samples of undergraduate writing graded by trained essay readers

[2] GPA in undergraduate courses that required a "considerable" amount of writing

[3] Comparison of writing with peers in major field of study (well below average to well above average)

[4] Reported success in college courses (not at all successful to extremely successful) with various kinds of writing (e.g., personal writing, persuasive writing, analysis/criticism, essay exams, etc.)

[5] Reported success (not at all to extremely) with various writing processes (e.g., organizing ideas and revising)

[6] Grades given by professors to the undergraduate writing samples evaluated in this study

[7] Reported accomplishments in writing (e.g., publishing a letter to the editor, writing technical manuals or other instructional material, authoring or co-authoring an article published in a scholarly journal, etc.)

#Median correlation over two individual readers

Note. Correlations are based on ns of 721 to 890. With n = 800, correlations of approximately .07 are significant at the .05 level; correlations of about .09 are significant at the .01 level. Between-correlation differences (*e-rater* versus two humans) of approximately .05 and .07 are significant at the .05 and .01 levels, respectively, using a test for differences between correlated correlations. All tests were two-tailed.

essays), the standard deviations are slightly larger for human reader scores than for *e-rater* scores for both issue (1.1 vs. 0.9) and argument essays (1.0 vs. 0.8). An examination of score distributions (Appendix A) confirmed that, as has been known, *e-rater* tends to award fewer scores that are either very high or very low, when compared to those assigned by human readers. Had we scaled *e-rater* scores to have the same variation as scores from human readers, the differences between *e-rater*'s and human readers' correlations with the nontest criteria would, predictably, have been somewhat smaller.

Table 2 displays the mean and standard deviation by level of performance on the GRE Writing Assessment for each of the most predictable nontest indicators. Also shown, by level of performance on the GRE Writing Assessment, are the percentages of participants at selected levels within indicators (for example, the proportion with GPAs of A+ in writing courses and the percentage whose writing samples received the highest scores of 5 or 6). Figures 1-9 graphically depict the data given in Table 2. As can be seen, mean standing for each of the nontest indicators generally increases with each higher level of performance on the GRE Writing Assessment. The progression is relatively similar regardless of the method of scoring used. It is also clear from these graphs, however, that inconsistencies between human and *e-rater* scores are most prominent at the upper and lower ends of the GRE Writing Assessment scale.

Discussion

Several research efforts have demonstrated a strong relationship between human and automated scores, thus implying the validity of computer-generated scores. This implication, however, is based on the assumption that scores assigned by human readers are themselves valid, and that they can, therefore, serve as the "gold standard" against which to validate automated scoring. The view taken here is that by using this standard, most previous studies have provided only an indirect test of the validity of automated scoring. Although unlikely perhaps, it is possible that human and automated scores could both be based, at least in part, on features that are not entirely relevant to good writing. Rather than attesting to the validity of automated scoring, substantial correlations between human and automated scores might instead reflect the sharing of largely irrelevant variation. For early automated systems, one such feature was, possibly, essay length -- that is, the number of total words in an essay. (*E-rater* does not use this feature.) Shown to relate strongly to human readers' evaluations, essay length was an influential feature in some systems (Petersen, 1997), even though its relevance to good writing has been questioned (Arden Albee, personal communication, January 11, 1999).[4]

Table 2

Performance on Nontest Indicators of Writing by GRE Writing Assessment Score Level for Human Scoring, *E-Rater* Scoring, and Both Combined

		GRE Writing Assessment Score Level					
Nontest indicators	Scoring method	1.5	2.5	3.25	4.0	4.75	5.5
				Mean (SD)			
Writing sample X	Two human readers (H)	3.3 (1.0)	3.4 (1.3)	3.7 (1.1)	4.0 (1.0)	4.4 (1.0)	4.7 (1.0)
	E-rater (E)	3.4 (0.9)	3.5 (1.2)	3.7 (1.1)	4.1 (1.1)	4.3 (1.0)	4.1 (1.1)
	Both H & E	3.2 (1.1)	3.4 (1.2)	3.7 (1.1)	4.1 (1.0)	4.4 (0.9)	4.7 (1.2)
Writing sample Y	Two human readers (H)	2.7 (1.1)	3.2 (1.0)	3.5 (1.1)	3.7 (1.0)	4.0 (1.0)	4.2 (1.1)
	E-rater (E)	3.3 (1.0)	3.2 (1.0)	3.5 (1.1)	3.8 (1.1)	3.9 (0.9)	3.7 (1.3)
	Both H & E	3.0 (1.1)	3.1 (1.1)	3.5 (1.1)	3.8 (1.0)	3.9 (1.0)	4.1 (1.3)
GPA in writing courses	Two human readers (H)	4.9 (1.0)	5.3 (1.1)	5.5 (1.1)	5.9 (1.0)	6.1 (1.0)	6.6 (0.7)
	E-rater (E)	4.9 (1.2)	5.2 (1.2)	5.6 (1.1)	5.9 (1.0)	6.2 (0.9)	6.1 (0.9)
	Both H & E	4.6 (0.8)	5.3 (1.2)	5.6 (1.1)	5.9 (1.0)	6.2 (0.9)	6.5 (0.8)
Comparison with peers	Two human readers (H)	3.3 (0.7)	3.6 (0.8)	3.7 (0.8)	3.9 (0.7)	4.1 (0.7)	4.2 (0.7)
	E-rater (E)	3.6 (0.9)	3.5 (0.8)	3.8 (0.8)	3.9 (0.8)	4.0 (0.8)	4.2 (0.7)
	Both H & E	3.1 (0.6)	3.6 (0.7)	3.7 (0.8)	3.9 (0.7)	4.1 (0.7)	4.3 (0.8)
				Percent top scores (5 or 6)			
Writing sample X	Two human readers (H)	10	23	22	31	42	69
	E-rater (E)	13	22	24	38	39	29
	Both H & E	14	20	23	34	41	67
Writing sample Y	Two human readers (H)	5	8	16	23	27	37
	E-rater (E)	6	8	18	24	25	32
	Both H & E	5	10	17	25	26	42
				Percent A+			
GPA in writing courses	Two human readers (H)	5	15	21	30	42	67
	E-rater (E)	13	17	23	30	47	45
	Both H & E	0	15	21	33	48	68
				Percent above average			
Comparison with peers	Two human readers (H)	30	58	64	76	81	86
	E-rater (E)	53	49	67	75	76	95
	Both H & E	27	56	66	75	82	90
N human/N *e-rater*/N both		43/17/23	133/83/144	283/471/340	261/169/256	143/136/112	38/20/20

Note. Ns range from 835 to 888 for the various indicators.

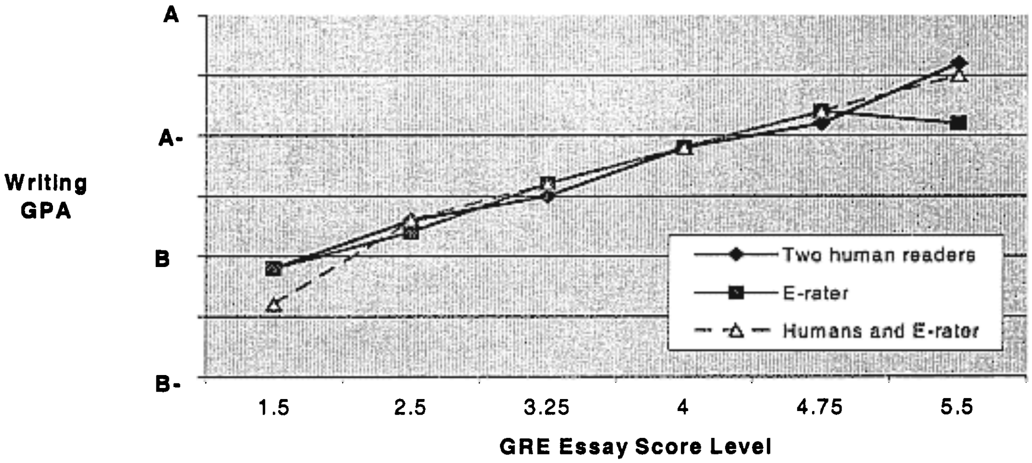

Figure 1: Relationship Between GRE Writing Assessment Scores and Undergraduate GPAs in Courses Requiring Writing, by Method of Scoring

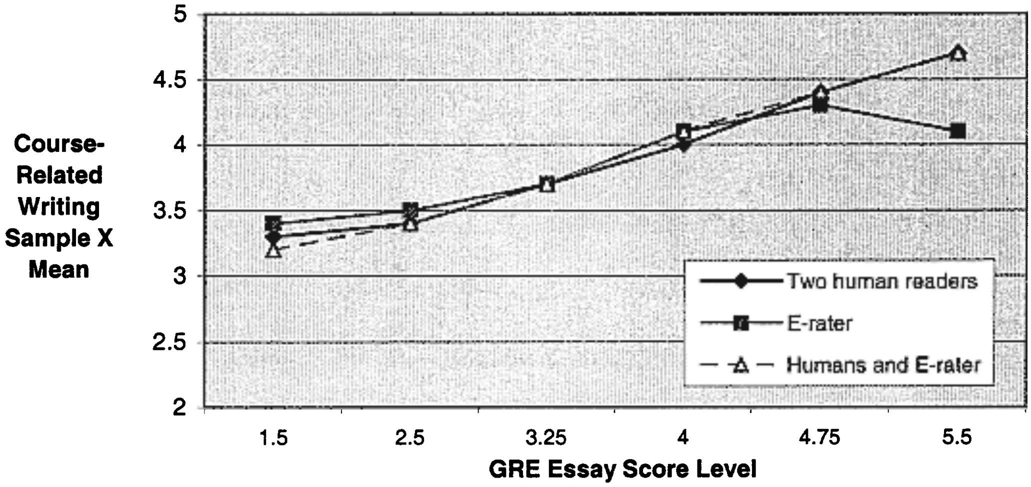

Figure 2: Relationship Between GRE Writing Assessment Scores and Course-Related Writing Samples, by Method of Scoring

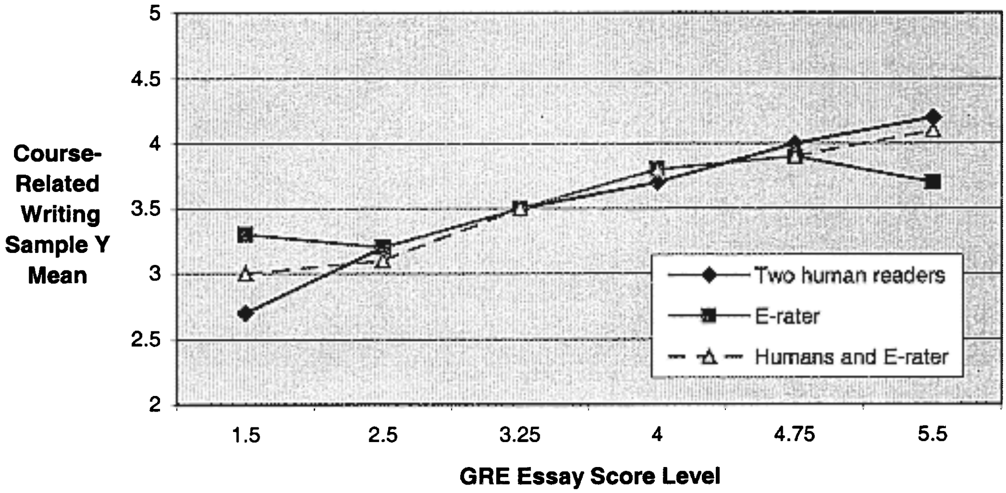

Figure 3: Relationship Between GRE Writing Assessment Scores and Course-Related Writing Samples, by Method of Scoring

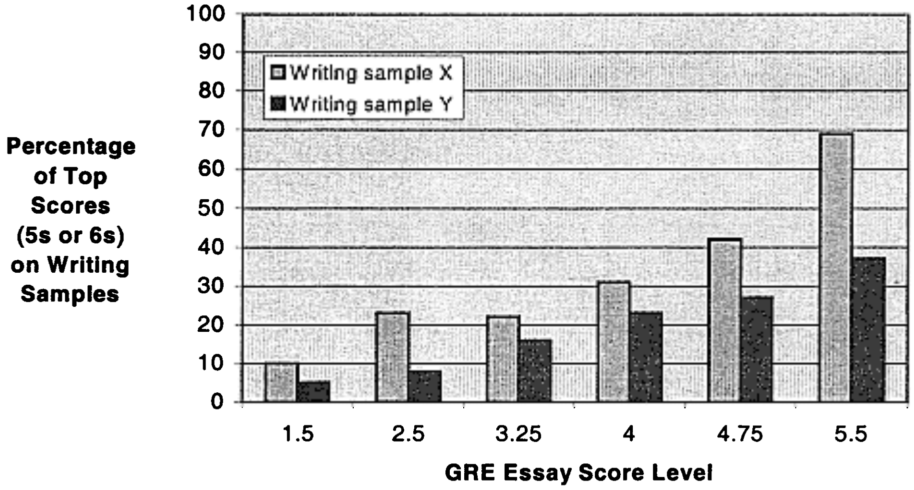

Figure 4: Performance on Course-Related Writing Samples by GRE Writing Assessment Level, as Graded by Two Human Readers

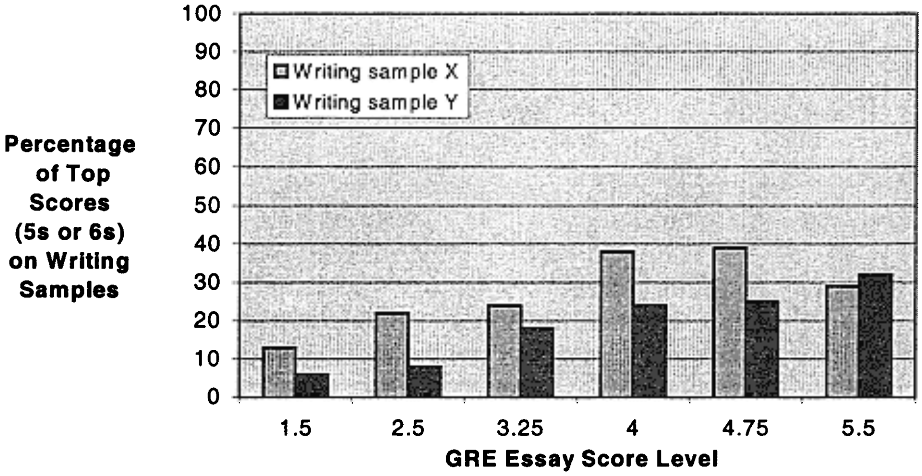

Figure 5: Performance on Course-Related Writing Samples by GRE Writing Assessment Level, as Graded by *E-Rater*

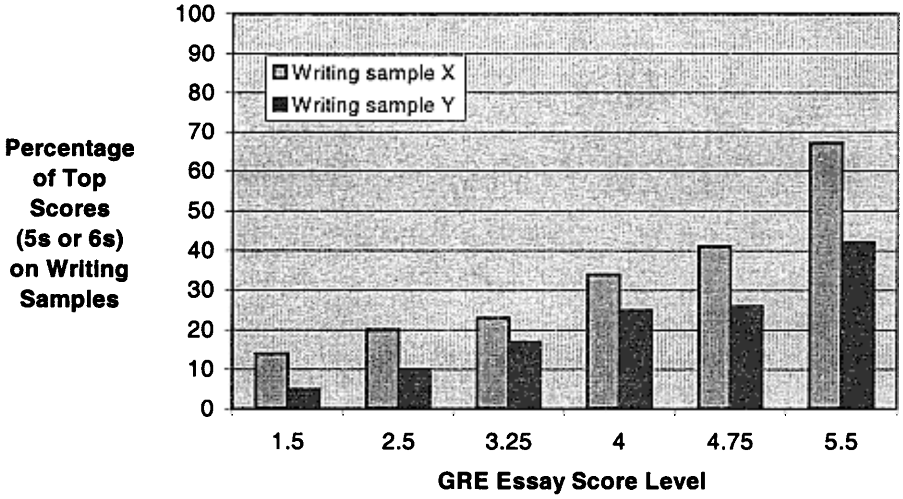

Figure 6: Performance on Course-Related Writing Samples by GRE Writing Assessment Level, as Graded by Two Humans and *E-Rater*

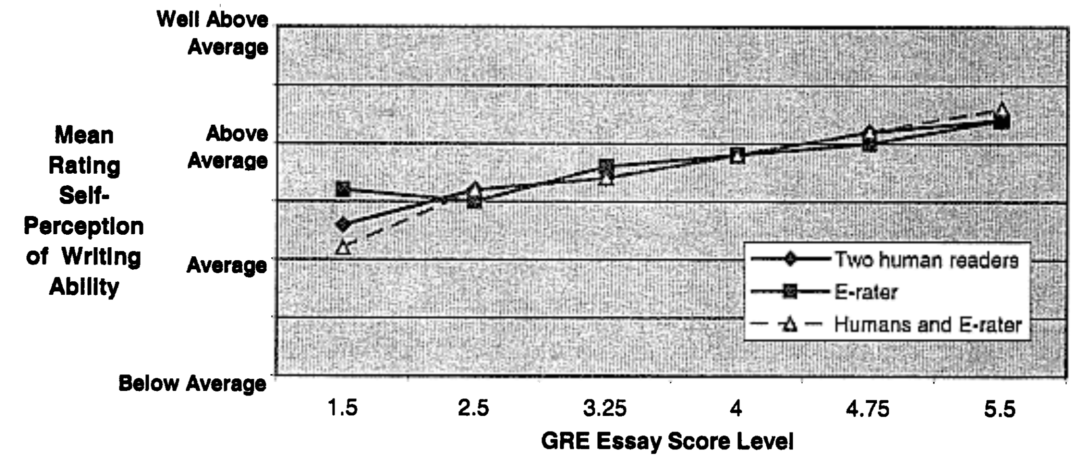

Figure 7: Self-Reported Ratings Versus GRE Writing Assessment Scores

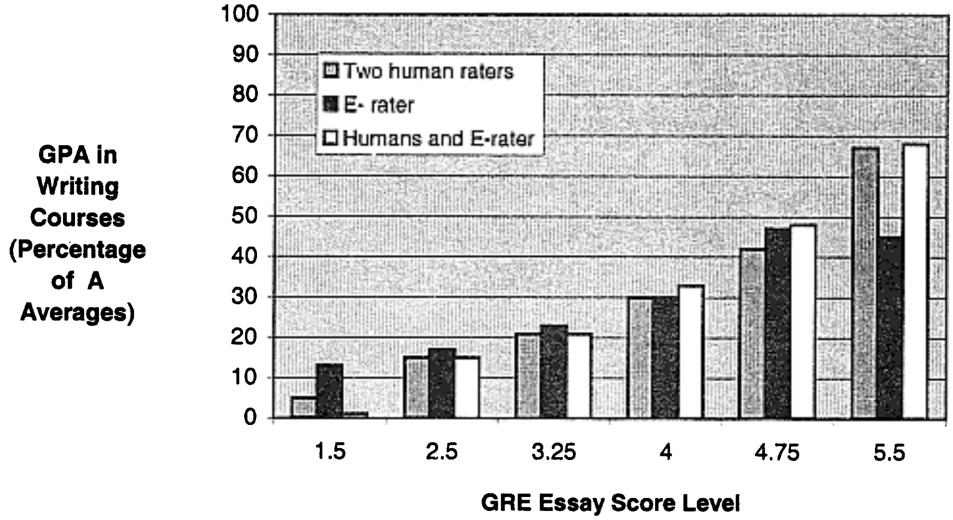

Figure 8: Self-Comparison with Peers by Performance on GRE Essays, by Method of Scoring

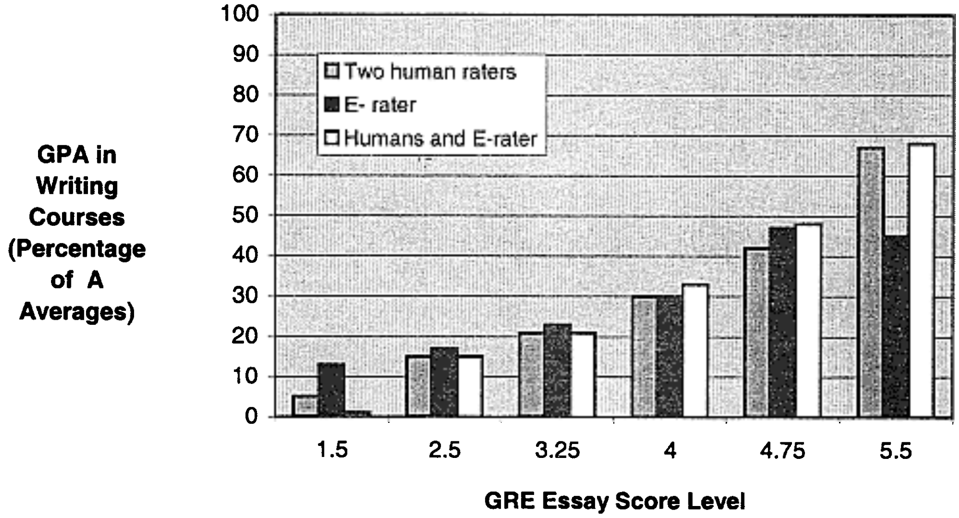

Figure 9: GPA in Writing Courses by Performance on GRE Essays, by Method of Scoring

The results of the study reported here suggest that automated scores, at least those generated from one particular scoring system (*e-rater*), do relate modestly to a number of indicators of writing skill. Moreover, they do so in a fashion that is reasonably similar to the relationships of these indicators to scores assigned by human readers. The relationships noted for automated scores were, however, somewhat weaker than those detected for scores from human readers. The weaker relationships noted for automated scores may be due to several factors, including:

- *e-rater*'s failure to focus (to the same degree as human readers) on the features of writing that are reflected in the various nontest indicators
- *e-rater*'s tendency to assign less variable (i.e., more intermediate and fewer extreme) scores than human readers
- *e-rater*'s reliance on specific, individual features, instead of on more general (and possibly more stable) dimensions or factors
- our use (in this study) of less-than-optimal data on which to train *e-rater*

Continuing research on *e-rater* may help to overcome the limitations imposed by each of these factors. The correlations of *e-rater* scores with the validation criteria we have considered might, for example, be raised by:

- supplementing *e-rater* models with additional essay features
- exploring the dimensionality of *e-rater* features in order to derive clusters of features (factors) that are more stable/reliable (and hence more predictive) than individual features
- employing fully adequate data[5] on which to specify *e-rater* scoring models
- considering interactions among the features on which *e-rater* bases its scores
- investigating alternative ways of scaling *e-rater* scores so as to improve their discrimination
- using multiple models (or different models at different score levels), for example, employing different sets of features to predict high scores than low ones

We suggest also that another avenue worthy of further consideration is the use of general *e-rater* models, as opposed to prompt-specific ones, as our use of general models in this study did not seem to degrade *e-rater*'s performance as much as we anticipated it might. On one hand, the use of a single, general model could render any writing assessment more susceptible to coaching or to other attempts to outwit *e-rater*. However, this disadvantage may be outweighed from the viewpoint of test validity and generalizability, as the use of a common model for more than a single prompt would signify that that the same features are valued regardless of the particular prompt.

Additional research is presently continuing on alternative methods of selecting and combining the essay features that *e-rater* extracts. For example, besides the use of ordinary least squares regression analysis, consideration is being given to non-linear regression methods, alternative classification techniques (e.g., tree-based regression), and procedures that employ neural nets. Each of these techniques could increase the validity of *e-rater* scores.

For now, however, we must settle for the evidence generated here, which suggests the potential of automated scores as (valid) indicators of prospective graduate students' writing skills, especially when they can be combined with scores provided by at least one human reader. It is our hope that research will continue on the development and improvement of automated scoring methods, and that improved versions of systems like *e-rater* will undergo evaluation of the kind we have undertaken here, as well as other kinds of scrutiny. Further research to probe the limits of automated scoring seems desirable. A specific focus might be identifying the kinds of essays or writing styles that systems like *e-rater* may undervalue, as well as the ones that they may overrate. Further evidence of this sort is necessary, we believe, but probably not sufficient, if automated scores are to be accepted by the public and by the writing assessment community.[6]

References

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays (ETS Research Report RR-98-15). Princeton, NJ: Educational Testing Service.

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. In Computer-Mediated Language Assessment and Evaluation of Natural Language Processing. Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, College Park , Maryland.

Daigon, A. (1966). Computer grading of English composition. English Journal, 55, 46-52.

DeLoughry, T. J. (1995, October 20). Duke professor pushes concept of grading essays by computer. Chronicle of Higher Education, 42, A24-25.

Kaplan, R. M., Wolff, S., Burstein, J. C., Lu, C., Rock, D., & Kaplan, B. (1998). Scoring essays automatically using surface features (GRE Report No. 94-21 and ETS Research Report No. RR-98-30). Princeton, NJ: Educational Testing Service.

Labi, N. (1999, February 27). When computers do the grading. Time Magazine, 153, p. 57.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th Annual Conference of the Cognitive Science Society (pp. 412-417). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological Bulletin, 111, 172-175.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Mitchell, J. S. (1998, May 27). Commentary: SATs don't get you in. Education Week on the Web [On-line serial]. Available: http://www.edweek.org/ew/current/3mitch.h17

Murray, B. (1998, August). The latest techno tool: Essay-grading computers. APA Monitor, 29, p. 43.

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

Page, E. B. (1968). Analyzing student essays by computer. International Review of Education, 14, 210-225.

Page, E. B. & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. Phi Delta Kappan, 76, 561-565.

Petersen, N. S. (1997, March). Automated scoring of writing essays: Can such scores be valid? Paper presented at the annual meeting of the National Council on Education, Chicago, IL.

Powers, D. E., Fowles, M. E., & Welsh, C. K. (1999). Further validation of a writing assessment for graduate admissions (GRE Board Report No. 96-13R and ETS Research Report No. RR-99-18). Princeton, NJ: Educational Testing Service.

Schaeffer, G. A., Fowles, M. A., & Briel, J. B. (in press). Assessment of the psychometric characteristics of the GRE writing test (GRE Board Report No. 96-11). Princeton, NJ: Educational Testing Service.

Schwartz, A. E. (1998, April 26). Graded by machine. The Washington Post, p. C07.

Scott, J. (1999, January 31). Looking for the tidy mind, alas. The New York Times, Sec. 4, p. 2

Appendix A

Crosstabulation of *E-Rater* and Human Scores

Crosstabulation of *E-Rater* and Human Scores

	Human Reader Score Level								
E-Rater Score Level	4 – 8	9 – 11	12 – 14	15 – 17	18 – 20	21 – 24	Total	Percent exact agreement	Percent exact or adjacent agreement
21 – 24	0	0	0	1	9	**10**	20	50	95
18 – 20	0	0	8	37	**70**	21	136	51	94
15 – 17	1	8	27	**84**	44	5	169	50	92
12 - 14	15	88	**215**	136	17	2	473	45	93
9 – 11	17	**32**	31	2	2	0	84	38	95
4 – 8	**10**	5	2	0	0	0	17	59	88
Total	43	133	283	260	142	38	899	47	93
Percent exact agreement	23	24	76	32	49	26	47		
Percent exact or adjacent agreement	63	94	96	99	87	82	93		

Note. Entries for human readers are the sum of two readers each for the argument and issue essays; for *e-rater*, entries are twice the sum of *e-rater* scores for the two essays.

Endnotes

[1] Professor Art Young (personal communication, January 11, 2000) "chided" us mildly on our use of the term "human readers" throughout this report. It struck Art as odd that we felt the need to describe people who read and evaluate essays as being "human," so as to distinguish them from computer or machine readers. As Art pointed out, computers aren't really readers at all, at least in the ordinary sense of the word. We apologize, therefore (but only slightly!), for our use of this term, which in retrospect seems even to us to be "computer-centric."

[2] A sample of lower quality was requested because we have found in previous research that students tend to submit the best samples that they can muster, not typical samples. Our request was intended to result in a better representation of students' course-related writing.

[3] Correlation coefficients would, of course, have been significantly higher if we had corrected for attenuation due to criterion unreliability. We have chosen, however, to report uncorrected correlations when making comparisons between relations for *e-rater* and human scores.

[4] Essay length, as reflected by the total words in an essay, is a simple enough concept. However, the import of the number of words in an essay is open to debate. On one hand, essay length may represent the in-depth development of cogent ideas; on the other hand, it may merely reflect extraneous palaver.

[5] We note here that for two of the 40 essay prompts considered in this study, we did, as a result of a separate data collection effort, have a sufficient number of essays on which to train *e-rater*. Consequently, it made sense to see if *e-rater* scores for these two prompts -- one argument prompt and one issue prompt -- bore stronger correlations with the nontest indicators than did scores from other prompts for which less training data were available. This examination was, however, inconclusive: The correlations with nontest indicators were somewhat higher for these prompts for some, but not all, of the indicators. However, the correlations of human scores with the indicators were also higher for these prompts for several, but again not all, of the indicators. Thus, there was no clear evidence that the lack of adequate training data dampened the correlations between *e-rater* scores and nontest indicators. Details about these results are available from the authors.

[6] As a teacher of writing, Art Young (personal communication, January 11, 2000), for one, remains unconvinced that the kind of evidence we have presented will be regarded as compelling by the public and the writing community members whom he knows. In Art's view, a critical shortcoming of automated scoring methods is their inability to assess such aspects of writing as a writer's ability to connect with and communicate to human readers. Moreover, the use of automated scoring may contribute further to the prevalence of "test-taking writing," where the aim is to convince an evaluator of one's ability to write in a prescribed manner (rather than in a way that communicates with readers, challenges them to think, or inspires them to act). In this view, automated scoring may, therefore, further devalue writing as an act of meaningful communication. These concerns are, of course, quite legitimate ones that, although beyond the scope of the study reported here, must be addressed.