GRE®
RESEARCH

Stumping *E-Rater*: Challenging the Validity of Automated Essay Scoring

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

March 2001

GRE Board Professional Report No. 98-08bP

ETS Research Report 01-03

ETS® Educational Testing Service
Princeton, NJ 08541

Stumping *E-Rater*: Challenging the Validity of Automated Essay Scoring

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

GRE No. 98-08bP

March 2001

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.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Educational Testing Service
Princeton, NJ 08541

Abstract

For this study, various writing experts were invited to "challenge" *e-rater* -- an automated essay scorer that relies on natural language processing techniques -- by composing essays in response to Graduate Record Examinations (GRE®) Writing Assessment prompts with the intention of undermining its scoring capability. Specifically, using detailed information about *e-rater*'s approach to essay scoring, writers tried to "trick" the computer-based system into assigning scores that were higher or lower than deserved. *E-rater's* automated scores on these “problem essays” were compared with scores given by two trained, human readers, and the difference between the scores constituted the standard for judging the extent to which *e-rater* was fooled. Challengers were differentially successful in writing problematic essays. Expert writers were more successful in tricking *e-rater* into assigning scores that were too high than in duping *e-rater* into awarding scores that were too low. The study provides information on ways in which *e-rater*, and perhaps other automated essay scoring systems, may fail to provide accurate evaluations, if used as the sole method of scoring in high-stakes assessments. The results suggest possible avenues for improving automated scoring methods.

Keywords: Writing assessment, Graduate Record Examinations (GRE), validity, automated scoring, essay scoring, *e-rater*

Acknowledgements

The authors extend their sincere thanks to the many people who responded to our request by writing essays designed to challenge *e-rater*. We are very grateful to Ruth Yoder, who tracked and processed the responses that we received and who oversaw various administrative aspects of the research. Finally, we are appreciative of the thoughtful reviews of an earlier draft that were provided by Leona Aiken, Isaac Bejar, Randy Bennett, Carol Dwyer, Joseph Jones, Larry Stricker, Rich Swartz, Cyndi Welsh, and Art Young. We are grateful to the Graduate Records Examination (GRE) Research Committee for its support of this research.

Table of Contents

List of Tables

Introduction

Complex performance assessments that typically require test takers to perform or produce (instead of to recognize or select) are becoming increasingly popular (Aschbacher, 1991). Although such constructed-response measures offer distinct advantages over traditional multiple-choice measures (see Bennett & Ward, 1993, for example), they usually have certain limitations as well. One unmistakable drawback, for instance, is that, compared to machine-scored multiple-choice questions, constructed responses ordinarily require a much more labor-intensive scoring effort, especially for essay measures of writing skill.

One hope for reducing the cost of essay scoring comes from longstanding efforts to develop computer programs that can, by modeling trained essay readers, evaluate essays automatically. The prospect of computer-based essay scoring was in fact envisioned more than 30 years ago (Page, 1966, 1968), and although there is clearly still much to be accomplished, computer-assisted evaluation of student essays is now becoming a feasible option (Page & Petersen, 1995). However, in order to improve automated scoring and increase its acceptance, researchers need to better understand the kinds of challenges presented by automated scoring. The objective of this study was to anticipate one such challenge and to assess the gravity of the threat it may pose.

Perceptions of Automated Scoring

While there is clearly considerable skepticism about the prospects of automated essay scoring (Wresch, 1993), some reactions have been decidedly positive. In addition to being cost effective, computerized scoring is unfailingly consistent, highly objective, and entirely impartial (Schwartz, 1998; Weinstein, 1998). To some, however, the notion of computer-assisted evaluation is incompatible with current conceptions of writing proficiency, which stress the writer's ability to communicate or "connect with" a designated audience (Art Young, personal communication, January 11, 2000).

Unlike human readers, computers are incapable, according to some critics, of distinguishing exceptional, inspirational essays from those that, while technically correct, are clearly quite ordinary (DeLoughry, 1995; Mitchell, 1998). This perceived shortcoming stems, it seems, from the presumption that automated scoring emphasizes linguistic rules and grammatical conventions at the expense of less tangible qualities, such as clarity and coherence. According to this view, computers may be able to *analyze* writing for the presence or absence of certain words or structures (and to use this analysis to mimic trained readers), but they cannot really *understand* or *appreciate* a writer's message in the same sense that human readers can. Even the developers of automated systems readily acknowledge that -- although such systems can be useful tools in instruction and assessment -- they cannot replace effective writing instructors (Rich Swartz, personal communication, November 27, 2000).

Prospects

Some observers (Breland, 1996, for example) are at least somewhat sanguine about a rapprochement between the proponents and the critics of computer-aided evaluation of writing. If any such reconciliation is possible, it depends, we believe, on several factors. First, in order to dispel misconceptions and promote understanding, the techniques that underlie current, automated scoring methods must be properly elucidated and clearly communicated. To some extent, the criticism of current methods may be based, at least in part, on reactions to earlier, outdated procedures that tended to rely heavily on the evaluation of surface features, such as the number of words in an essay (Kaplan, Wolff, Burstein, Lu, Rock, & Kaplan, 1998). Understandably, such primitive methods often encountered considerable resistance from the writing community. Compared with these earlier counterparts, however, more recently developed procedures rely less on surface features and more on deeper structures of text. In addition, the variables underlying the newer models are based more strongly on the qualities of writing on which human readers are trained to focus (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998).

Second, once adequately understood, automated methods must also undergo thorough evaluation by the harshest critics and most skeptical onlookers. This kind of scrutiny could help to reveal whether automated scoring is unduly susceptible to influences that bear little if any relevance to the construct of writing ability. Surviving such challenges (that is, by discounting the plausibility of the alternative interpretations that they pose) is in fact a hallmark of modern test validation (Kane, 1992; Messick, 1989, 1995).

Purpose

For automated essay scoring, one specific threat to validity is the extent to which it may improperly reward or, conversely, unfairly penalize certain features of examinees' writing. The goal of this study was to assess the plausibility of this "double-edged" threat to the validity of automated scoring. Both construct irrelevance and construct underrepresentation underlie the possibility, as follows:

- automated methods may be unduly influenced by *extraneous* features of examinees' writing (and therefore "tricked" into awarding higher-than-deserved scores)
- automated methods may fail to recognize features that are clearly *relevant* to good writing (and therefore be disposed to awarding lower-than-deserved scores)

Our general aim was to determine the extent to which either scenario poses a threat to the validity of automated essay scoring. Consistent with Perfetti's (1998) suggestion, we sought to distinguish between system weaknesses that are easily corrected and those that are potentially more serious because they result from fundamental differences in the ways that humans and computers process and analyze language. Hence, the specific questions of interest were:

- Do certain writing approaches or test-taking strategies give rise to large discrepancies -- either positive or negative -- between automated scores and scores awarded by readers trained on established scoring guidelines?
- What do these discrepancies suggest about the validity of inferences based on computer-generated scores?

To investigate these questions, various writing experts were invited to compose essays in response to Graduate Record Examinations (GRE$^{®}$) Writing Assessment prompts with the intention of tricking one automated scoring system -- *e-rater* -- into awarding scores that were either higher or lower than deserved.

Method

Materials/Instruments

The GRE Writing Assessment. The GRE Writing Assessment is designed to measure prospective graduate students' 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 complementary writing tasks: a 45-minute task that requires examinees to present their perspective on an issue, and a 30-minute task that requires them to analyze an argument. The prompt for the issue task states an opinion on a topic of general interest and asks writers to address the topic from any perspective they wish, providing relevant reasons and examples to explain and support their views. The prompt for the argument task presents a written argument and requires test takers to critique it, discussing how well reasoned they find the argument. Test takers are asked not to agree or disagree with the position expressed in the argument, but rather to evaluate and discuss its logical soundness. At least two trained, certified GRE readers evaluate each essay. (For more detail about the GRE Writing Assessment and how it is scored, see Powers, Fowles, and Welsh, 1999, and visit http://www.gre.org/twotasks.html.)

E-rater. The automated scoring method studied here was *e-rater,* a system that is being developed to evaluate test takers' responses to various kinds of essay tasks and prompts. In brief, *e-rater* uses natural language processing techniques to model the performance of human evaluators. For each essay prompt, assessors select a sample of essays that have been previously scored by at least two human readers and that represent the complete range of possible scores. The system then analyzes these essays for certain features and "trains" itself to recognize similar features in other essays.

Designed to model human readers (in this case, the *outcomes* of their evaluations, not necessarily the *processes* they use), *e-rater* first uses several subroutines to extract a wide variety of features from test takers' essays. A subset of the most predictive features, usually about eight to 12 in number, is retained and used in combination to predict the scores that would be assigned by human readers, based on the features of the essays in *e-rater's* training sample. A different model is specified for each essay prompt, and the regression weights for the model's predictors are applied to each new essay in order to estimate a score, which is rounded to the nearest integer from 0 to 6 -- the same 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 discourse features that occur throughout extended text), and *content* (indicated by prompt-specific vocabulary). Structural characteristics include "syntactic variety" -- that is, the use of various clause and verb structures. Organization characteristics include a variety of discourse features (i.e., characteristics that are associated with the orderly presentation of ideas) as well as linguistic cues (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 examines an essay's *content* -- its prompt-specific vocabulary -- both argument-by-argument and for the essay as a whole. The expectation is that words used in better essays will bear a greater resemblance to those used in other good essays than to those used in weak ones, and that the vocabulary of weak essays will more closely resemble words in other weak essays. Programmed on these assumptions, *e-rater* assigns a number to each essay on the basis of the similarity of its vocabulary to that of other previously scored essays. (More detail about how *e-rater* works is available in several reports -- such as Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock, & Wolff, 1998 -- that can be downloaded from the following Web site: http://www.ets.org/research/erater.html.)

Study Participants

Several distinct classes of study participants were asked to write essays they believed would challenge *e-rater*. These challengers included:

- staff from the Educational Testing Service (ETS®) Assessment Division -- in particular, those involved in the development of the GRE Writing Assessment

- researchers (e.g., computational linguists and cognitive scientists)
- specialists in artificial intelligence and text classification
- members of the writing community, especially those known to be critical of automated scoring methods
- specialists in English as a second language (ESL) and nonnative speakers of English
- others known, by virtue of their publications or through personal contact with us, to have an interest in the automated scoring of essays

Procedures

Potential participants were contacted by several means and invited to participate in a study whose purpose was to challenge *e-rater*. Two major listserves, one for computational linguists and the other for language testers, provided the primary means of extending our invitation. Other parties known to us to have interest in automated scoring were also invited to participate. Further, invitees were encouraged to forward our invitation to others whom they felt might have the interest and expertise to challenge *e-rater*.

Our general approach was to invite writing experts and other interested parties to compose essays that they thought *e-rater* would find problematical and, as a result, would award scores that were either higher or lower than deserved, relative to the scores awarded by human readers. Participants were unpaid, but were promised feedback on the success of their efforts to fool *e-rater*, as well as a summary of the success of other study participants. The only monetary incentive to participate was a chance to win one of two $250 prizes by submitting an essay that was either greatly undervalued or greatly overrated by *e-rater*.

A total of four GRE Writing Assessment prompts -- two argument and two issue (see Appendix A) -- were identified for use in the study. Each of these prompts had survived earlier pretesting and was judged to be representative of the total pool of those available for the GRE Writing Assessment. In addition, an *e-rater* scoring model had been specified previously for each of these prompts. These models were deemed to be reasonably representative of the models that had been developed for other prompts.

On the basis of the relevant research conducted to date, a description of *e-rater* was prepared as a guide for our challengers. The description (see Appendix B) explains the general approach taken by *e-rater*, the specific techniques it uses, and the particular cue words on which it focuses. All participants received this description, along with information about the GRE Writing Assessment, copies of the scoring guides used by GRE readers, two essay prompts -- one issue and one argument -- on which they were to write their essays, and samples of previously scored responses to these prompts. Half of the writers randomly received one pair of prompts and half the other pair.

Study participants were asked to write two essays for each of the two prompts they received -- one that they predicted would elicit a score from *e-rater* that was higher than deserved and one that they thought would receive an *e-rater* score that was lower than deserved. For each essay submitted, writers were asked to explain the reasons for the discrepancies they predicted. All essays were then collected and scored both by *e-rater* and also by two trained human readers using the published holistic GRE scoring criteria. Readers who evaluated challengers' essays were aware that the essays were written specifically for the study, but they were blind to the specific intentions of the writers. Once essays were scored, the discrepancy between the *e-rater* score and the average score assigned by readers was computed for each essay.

Results

Description of the Sample

In total, 27 people responded to our challenge by writing one or more essays. The titles or positions held by these respondents, and an indication of their relevant experiences and interests, are shown in Table 1. Undergraduate and graduate students from language-related disciplines were strongly represented in the sample.

Table 1

Description of Study Participants

Title/Position	Relevant Experience/Interests
Emeritus professor of applied linguistics, University of Edinburgh	Forty years experience in language teaching, applied linguistics, and development of language tests
Professor of linguistics, director of studies of the Ph.D. program in language pedagogy, Budapest University	British Council advisor to the Hungarian Examination Reform Project; editor, Language Testing; co-editor, Cambridge Language Assessment Series
Academic coordinator, Language Center, Al Akhawayn University, Morocco	None specified
Undergraduate student (n = 6) in advanced course in natural language processing, University of Texas at Austin	None specified
Undergraduate student (n = 9) of ESL, Montclair State University, New Jersey	None specified
Writing assessment specialist (n = 3) at ETS	Extensive experience developing writing assessments for a variety of testing programs
Assistant or associate professor (n = 2) of computational linguistics, University of Rochester, University of Pennsylvania	Knowledge of statistical natural language processing
Doctoral candidate in text categorization, Portsmouth University	Interest in applications of categorization software
Doctoral candidate, Modern Language Center, University of Toronto	Interest in language assessment, especially adult learners of ESL; experience as an ESL teacher and researcher
Graduate student, Rensselaer Polytechnic Institute	Interest in logic-based artificial intelligence

Evaluation of Essays (Descriptive Statistics)

The 27 participants returned a total of 63 essays. The mean scores assigned to essays by first and second readers, respectively, were 3.22 (sd = 1.45) and 3.26 (sd = 1.54). Trained readers agreed exactly with one another 52% of the time, while *e-rater* agreed exactly with individual readers about 34% of the time. Readers agreed exactly or within one point of one another 92% of the time, while *e-rater* exhibited this level of agreement with individual readers approximately 65% of the time.

The product-moment correlation between readers was .82, while the correlations between *e-rater* and individual first and second readers were .42 and .37, respectively. The corresponding statistics for Cohen's kappa were .42 between readers, and .16 and .27 between *e-rater* and individual first and second readers, respectively. For agreement exactly or within one point, Cohen's kappa was .85 between readers, and .49 and .27 between *e-rater* and individual first and second readers, respectively. By any measure, therefore, *e-rater*'s agreement with trained readers was less than agreement between readers.[1] Typically, the rate of agreement (exact or within one point) computed in previous studies of *e-rater* has been 90-95 % between readers and *e-rater*, and the correlation has been in the mid .80s (Burstein, Kukich et al., 1998).

Summary and Accuracy of Predictions

Table 2 shows, by hypothesized and actual discrepancies, the means and standard deviations of scores awarded by *e-rater* and by human readers. A variety of information about the essays that were submitted can be extracted from this table. First, of the total 63 essays, nearly half (30) were expected to receive a higher score from *e-rater* than from human readers, and a somewhat smaller number (24) were predicted to receive a lower score from *e-rater*. For the remaining nine essays, no prediction was made. In actuality, a slight majority of the essays (39, or 59%) were assigned higher scores by *e-rater* when compared with the average of scores assigned by human readers. The remaining essays were about equally likely to receive either a lower score from *e-rater* (22%) or to get exactly the same score from readers and from *e-rater* (19%). In total, 36 of 54 predictions (67%) were in the correct direction. Another nine predictions (17%) were in the wrong direction. The remaining nine (17%) were also incorrect, in the sense that there were no discrepancies in these cases.

Table 2

Means and SDs for *E-Rater* and Readers by Hypothesized and Actual Discrepancies

Hypothesized		Actual direction of discrepancy							
direction of		(Score higher for)							
discrepancy		*E-rater*		Trained readers		No discrepancy		Total	
(Score higher for)		*E-rater*	Readers	*E-rater*	Readers	*E-rater*	Readers	*E-rater*	Readers
E-rater	N	26		3		1		30	
	M	4.27	2.60	4.33	5.67	3.00	3.00	4.23	2.92
	SD	1.19	0.92	0.47	0.24	0.00	0.00	1.15	1.26
Trained readers	N	6		10		8		24	
	M	4.33	2.83	3.70	4.80	3.75	3.75	3.88	3.96
	SD	1.11	1.28	0.64	0.87	1.79	1.79	1.27	1.55
Unspecified	N	5		1		3		9	
	M	3.40	2.40	3.00	3.50	2.33	2.33	3.00	2.50
	SD	0.49	0.58	0.00	0.00	0.47	0.47	0.67	0.62
Total	N	37		14		12		63	
	M	4.16	2.61	3.79	4.89	3.33	3.33	3.92	3.25
	SD	1.15	0.96	0.67	0.91	1.60	1.60	1.21	1.43

Table 3 provides the mean discrepancies and standard deviations for each combination of hypothesized and actual discrepancies. For the 26 essays that were correctly hypothesized to get a *higher* score from *e-rater* than from human readers, three were discrepant by 3 or more points, five by 2 to 3 points, 12 by 1 to 1.5 points, and six by .5 points. For the 10 essays that were correctly hypothesized to receive a *lower* score from *e-rater* than from readers, three were discrepant by 2 points, three by 1 point, and four by .5 point.

Table 3

Correspondence Between Hypothesized and Actual Discrepancies Between Scores Assigned by Human Readers and by *E-Rater*

Hypothesized direction of discrepancy (Score higher for)		Actual direction of discrepancy (Score higher for)			
		E-rater	Trained readers	No discrepancy	Total
E-rater	N	26	3	1	30
	M	+1.67	−1.33	0.0	
	SD	1.18	0.62	0.0	
Trained readers	N	6	10	8	24
	M	+1.50	−1.10	0.0	
	SD	0.96	0.62	0.0	
Unspecified	N	5	1	3	9
	M	+1.00	−0.5	0.0	
	SD	0.55	0.0	0.0	
Total		37	14	12	63

Note. Discrepancies were computed by subtracting the average of two reader scores from the *E-rater* score.

Thus, predictions that *e-rater* would award a higher score than human readers were borne out in 26 of 30 instances (87%), while predictions that *e-rater* would award a lower score than human readers were accurate less often (10 of 24, or 42% of the time). The mean discrepancies between *e-rater* and human reader scores were approximately equal in size, regardless of the direction of the prediction, and also regardless of whether or not the prediction was accurate. That is, when participants were wrong, discrepancies were often just as large (in the opposite direction) as when they were right.

Bases for Predictions

Appendix C lists the reasons writers offered when predicting that *e-rater* would give a higher than deserved score. Appendix D provides the same information for predictions that *e-rater* would award a lower score than deserved. These appendices also show the actual discrepancies between *e-rater* and reader scores.

The clear "winner" in our contest was an issue essay submitted by a professor of computational linguistics. His principal strategy was simply to write several paragraphs and to repeat them (37 times, in fact!). This strategy did indeed fool *e-rater*, resulting in a maximum discrepancy of 5 points. *E-rater* assigned the essay the highest possible score (6), while both study readers awarded it the lowest possible score (1). As we will discuss later, work is continuing on procedures for detecting essays of this nature.

The second most successful challenger employed a slight variation on the winning strategy, again basically repeating the same paragraph, but rewording the first sentence of each paragraph slightly, substituting a few key words, and reordering the subsequent sentences. Like the winning entry, this essay (also based on an issue prompt) received the top score (6) from *e-rater*, but in this case both human readers awarded it a score of 2.

The third most effective challenge, resulting in a discrepancy of 3.5, was written in response to an argument prompt. In this relatively long response, the writer attempted to stress the features that *e-rater* attends to. He varied sentence structure, used discourse cue words, and employed vocabulary that was related closely to the subject matter of the essay prompt. The essay deserved a low score, according to the writer, because it failed to provide any critical analysis, and instead merely agreed with various aspects of the argument. *E-rater* bestowed a top score of 6; trained GRE readers assigned scores of 2 and 3.

Three other essays were each scored 3 points higher by *e-rater* than by readers. Each of these essays -- two written to issue prompts and the other to an argument prompt -- attempted, alternatively, to write essays that “rambled,” “missed the point,” used faulty logic, or were “haphazard” in their progression, but used relevant content words, complex sentence structure, or other features valued by *e-rater*.

While it was clearly possible to write essays that *e-rater* scored too high, it appeared more difficult to write essays that tricked *e-rater* into giving lower-than-deserved scores. Only two issue essays received *e-rater* scores that were two points lower than reader scores. No other discrepancies in this direction were greater than one point. Both of the “winners” in this category made extensive use of metaphor and literary allusion, and avoided the use of *e-rater* clue words, instead using, as one writer characterized them, “more subtle transitions.”

Discussion

Much of the discussion that follows presumes that *e-rater*, or some variant of it, will play a role in evaluating examinee responses to the GRE Writing Assessment. Regardless of whether this assumption is warranted, it seems unequivocal that essays written in “bad faith” can reduce the correspondence between *e-rater* scores and those awarded by trained readers. Furthermore, this reduction may be dramatic -- to a level that is considerably lower than the agreement rates typically noted in the studies cited earlier that compared automated and reader-based essay scores.

Our results suggest also that it is easier to dupe *e-rater* into awarding undeservedly high scores than it is to write (either purposively or inadvertently) in a manner that results in scores that are too low. That the latter is somewhat more difficult is good news, we believe, as it suggests that examinees are relatively unlikely to receive *e-rater* scores that are lower than deserved because they submit essays that are unusually clever, highly creative, or unique in potentially counterproductive ways. The advice offered by one coaching school for the Graduate Management Admission Council’s Analytical Writing Assessment -- to “be a conformist” because the computer can’t appreciate individuality, humor, or poetic inspiration (http://800score.com/gmat-essay.html) -- seems less compelling in light of these results. Indeed, it would be unfortunate to discourage originality because of the current state of automated scoring. Further research would seem warranted, however, to ensure that examinees who write good, but unusual, essays are not beaten by the system.

On the other hand, suggestions (from the same website) intended to help examinees beat *e-rater* -- for example, by using phrases like “unwarranted assumption” and “fallacy of equivication [*sic*]” -- could have a more deleterious effect on the meaning of scores. Our results suggest that the gratuitous interjection of such phrases and cue words can affect *e-rater* evaluations to some degree, thus artificially inflating *e-rater* scores. It is also clear from our results, however, that fooling *e-rater* is apparently not as easy as some observers have supposed. In our study, a significant number of such attempts actually “backfired” in the sense that they resulted in essays that ran counter to writers’ predictions. In addition, although our study did not evaluate the following possibility, it seems plausible that, by indiscriminately using such phrases in order to fool *e-rater*, a writer might produce an essay that would be downgraded by trained readers in the "real world" of high-stakes testing. Like many so-called tricks of test-taking, strategies can sometimes be so complicated and convoluted (and therefore potentially counterproductive) that test takers are better advised to just "do their best" instead of trying to outwit the test maker, or in this case, the test scorer. This is clearly the best strategy when *e-rater* is used for instructional purposes, for as the developers of *e-rater* make clear to potential users, *e-rater* scores have meaning only if writers make genuine and legitimate attempts to respond to essay prompts.

In addition to their relevance for test takers, the study results also have implications for essay readers. These implications are especially relevant if scoring by human readers is paired with automated scoring, as is the current practice for scoring the GMAT writing assessment. In particular, the findings hold clues as to the features of test takers' writing on which readers should focus (or about which they should be especially diligent).

Finally, our findings point to certain aspects of writing that automated scoring may fail to notice or appreciate, or may reward unduly, and thus the results of this study provide some direction to improve *e-rater*. In particular, further research on how to identify excessively repetitive essays, as well as those that employ questionable logic, would be desirable. If such attempts prove unsuccessful, an alternative strategy, which is already being implemented, is to heighten readers' vigilance of features of writing that are problematical for *e-rater* and of test-taker efforts to beat the system. This strategy should further increase human readers' effectiveness in providing a "reality check" on *e-rater* scores.

Limitations

In some respects, our challenge to *e-rater* may have been less rigorous than desired. First, although we issued invitations to a variety of audiences, there is no assurance that we cast our net sufficiently wide. It seems certain that study participants did not represent all of the potentially successful challengers to *e-rater*. Second, because challengers were not selected to represent the test-taking population, they may not have been able to mimic the writing of typical GRE test takers (if this was in fact their intention). And third, we challenged only a relatively small number of essay prompts; while these prompts were representative of GRE Writing Assessment prompts in terms of difficulty and content, there is no assurance that they were representative with respect to their susceptibility to challenge.

In other ways, however, our challenge may have been somewhat more stringent than necessary. First of all, underlying the study was the implicit and unwarranted assumption that *e-rater* will be used *by itself* for high-stakes testing. Second, we allowed test developers -- challengers with "inside knowledge" -- to test *e-rater*. Moreover, in writing their challenges, participants were allowed some advantages that are not normally available to test takers. For example, challengers were not restricted to the time limits that are imposed on test takers. As we suggest below, future challenges might take other forms and involve additional categories of challengers.

A final possible limitation concerns the external validity of the study methods. For instance, the GRE readers who evaluated challengers' essays were aware that the essays were written specifically for our study. Furthermore, they noted that essays sometimes seemed unusual when compared to those normally written by test takers. Whether or not readers' would have been as sensitive to unusual essays, or would have provided similar evaluations for essays written under operational testing conditions, remains an open question.

Further Research

Perhaps the greatest value of the current study lies in what it suggests for future studies of *e-rater*. For example, with respect to writing effective challenges, some of our study challengers were consistently better than others were. In particular, ETS staff who have been involved in the development of writing assessments (and who have as a result evaluated hundreds of examinee essays) proved especially capable with regard to writing successful challenges. (Several admitted, however, that the task was not an easy one.) In future studies of *e-rater*'s susceptibility, it will therefore be important to continue the involvement of these staff members, even though in one sense the use of these insiders may seem an unfair test of *e-rater*. If, however, *e-rater* can withstand the vigorous challenges that these insiders can provide, then it should be better able to endure threats from other quarters as well. Of course, it will be important also to continue the participation of other interested parties.

With respect to further research, it may also be useful to consider alternative study designs. During the course of the study, several alternative challenge designs occurred to us. For instance, one possibility (which might have been less burdensome for study participants) would be to provide examples of essays at each score level and to ask challengers to modify these existing essays, rather than to compose entirely new ones. In this case, some modifications should be written to affect *e-rater* scores although, according to writing experts, they would have *no* effect on the quality of essays. Conversely, other changes should be predicted to have *no* impact on *e-rater* scores, when in fact these changes *do*, in the judgment of experts, have an effect on writing quality.

Yet another line of research would entail developing a package of successful challenge strategies and sharing these strategies with prospective GRE test takers. Performances on the GRE Writing Assessment under standard conditions would be compared with those obtained when examinees try to outwit *e-rater*. This focus is slightly different from that used in the current study, inasmuch as it would involve actual GRE test takers -- not experts in writing, computational linguistics, and so on. It would also focus exclusively on strategies presumed to result in higher scores than deserved.

An additional line of inquiry would involve monitoring examinee performance on the GMAT Analytical Writing Assessment, which employs automated scoring. Essays with score discrepancies between readers and *e-rater* could be analyzed to detect possible reasons for discrepant scores. Comparisons could also be made between the rates and kinds of discrepancies for assessments that do and do not use automated scoring. The aim here would be to ascertain the extent to which examinees use different writing strategies because they know that automated scoring will or will not be used to evaluate their essays. The GRE Writing Assessment, which currently uses two trained readers, and the GMAT Analytical Writing Assessment, which currently uses *e-rater* and one trained reader, would be natural comparisons. Once a sufficient database of discrepancies has been assembled, a more formal statistical analysis could be undertaken to determine the essay characteristics that best account for discrepancies between human and automated scores.

Finally, the website that has been developed to demonstrate scoring by *e-rater* (http://www.ets.org/criterion) could be utilized in future research. Essays written in response to this demonstration could be systematically collected and subjected to scoring both by trained readers and by *e-rater*. This kind of demonstration system, which has also been implemented by Landauer and his colleagues at the University of Colorado (http://lsa.colorado.edu/LSA-grade-main.html) for one automated scoring system (Landauer, Laham, Rehder, & Schreiner, 1997; Landauer, Foltz, & Laham, 1998), would enable a wide variety of interested parties to challenge the system. Challengers could submit essays for any of several topics and obtain almost instantaneous feedback in the form of automated scores. This approach would provide a convenient mechanism for communicating about *e-rater*, for allowing skeptics and others to experiment with it, and for evaluating the kinds of essays that might prove problematic.

Conclusion

This study has helped, we believe, to clarify both the promise and the potential pitfalls of one specific system for automated essay scoring. One clear implication of our results is that this system -- *e-rater* -- is not yet ready to “go it alone.” At least for high-stakes testing, it should not be used without the help of trained human readers. Indeed, neither ETS nor its subsidiary, ETS Technologies, advocates the use of *e-rater* by itself for high-stakes assessment (Rich Swartz, personal communication, November 27, 2000).

To date, the strategy for developing *e-rater* has been to train it to apply a model based on essay characteristics that trained readers assess. Our results suggest that *e-rater* and trained essay readers may react very differently to some features of examinees’ essays. Thus, in its current form, human intervention is required to keep *e-rater* from seriously misscoring some essays.

On the other hand, *e-rater* may be especially capable of systematically attending to some characteristics of examinees’ essays -- for example, certain details that trained readers are not expected to process fully when making global, impressionistic ratings, even with very deliberate reading. An option, therefore, might be to focus trained readers more narrowly on the qualities of writing that systems such as *e-rater* may not appreciate (or, alternatively, may overvalue), and to concentrate *e-rater’*s efforts on those characteristics that are regarded as indicators of good writing, regardless of how they are determined.

Postscript

Even as we conducted this study, other research was being carried out in an effort to improve *e-rater*. As a result, many of the successful challenges written for the study described here would probably not be successful today. For instance, some research has led to the development of techniques to detect essays that are unresponsive to essay prompts. In order to identify these so-called “off topic” essays, several filters have been devised -- some intended to detect essays that have very little overlap with the lexical content of the prompt, and others to spot writers’ tendencies to repeat substantive words. (Regarding the latter, the *lack* of repetition may signal that an essay is either very brief or else lacks development of a theme.)

When these new filters were applied to the 63 essays written for this study, a total of 16 essays were flagged for further inspection. Among these essays were both of the highly repetitive essays that had the largest positive discrepancies between *e-rater* and human readers’ scores, as well as one of the two essays that had the largest negative discrepancies. On the other hand, five essays that exhibited no score discrepancy were also flagged.

The effort to develop off-topic filters seems to hold promise. It is clear also however, even from this cursory analysis, that further research is needed. We hope that such research will be inspired by studies like the one we have reported here.

References

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4,* 275-288.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive assessment: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Breland, H. M. (1996). Computer-assisted writing assessment: The politics of science versus the humanities. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 249-256). New York: Modern Language Association of America.

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 No. 98-15). Princeton, NJ: Educational Testing Service.

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

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

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley & Sons.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

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

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.

Landis, J. D., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

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

Messick, S. (1995). Validity of psychological assessment. *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.

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.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes, 25*, 363-377.

Powers, D. E., Fowles, M. E., & Welsh, C. K. (1999). *Further validation of a writing assessment for graduate admissions* (GRE Board Research Report No. 96-13R and ETS Research Report No. 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. *Washington Post,* p. C07.

Weinstein, B. (1998, June 21). Software designed to grade essays. *Boston Globe Online* [On-line serial]. Available: http://www.Boston.com/dailyglobe/g…ware_designed_to_grade_essays.htm/

Wresch, W. (1993). The imminence of grading essays by computer – 25 years later. *Computers and Composition, 10*, 46-58.

Endnotes

[1] Although there are no precise standards for what constitutes sufficient interreader agreement, some guidelines have been offered. Fleiss (1981), for example, describes K ≥ .75 as reflecting "excellent" agreement and K ≤ .40 as signifying "poor" agreement. Similarly, Landis & Koch (1977) offered the following, more detailed categories:

Kappa	Strength of Agreement
< .00	Poor
.00 to .20	Slight
.21 to .40	Fair
.41 to .60	Moderate
.61 to .80	Substantial
.81 to 1.00	Almost perfect

Appendix A

Issue and Argument Prompts Used in the Study

ESSAY TOPICS

Issue: Absence of Choice

Present your perspective on the issue below, using relevant reasons and/or examples to support your views.

"The absence of choice is a circumstance that is very, very rare."

Argument: Worker Apathy

The following appeared in the editorial section of a corporate newsletter:

"The common notion that workers are generally apathetic about management issues is false, or at least outdated: a recently published survey indicates that 79 percent of the nearly 1,200 workers who responded to survey questionnaires expressed a high level of interest in the topics of corporate restructuring and redesign of benefits programs."

Discuss how well reasoned you find this argument.

Issue: Conformity/Success

Discuss the extent to which you agree or disagree with the opinion stated below. Support your views with reasons and/or examples from your own experience, observations, or reading.

"No one can possibly achieve any real and lasting success or 'get rich' in business by conforming to conventional practices or ways of thinking."

Argument: Roller Skating

Hospital statistics regarding people who go to the emergency room after roller-skating accidents indicate the need for more protective equipment. Within this group of people, 75 percent of those who had accidents in streets or parking lots were not wearing any protective clothing (helmets, knee pads, etc.) or any light-reflecting material (clip-on lights, glow-in-the-dark wrist pads, etc.). Clearly, these statistics indicate that by investing in high-quality protective gear and reflective equipment, roller skaters will greatly reduce their risk of being severely injured in an accident.

Discuss how well reasoned you find this argument.

Appendix B

Description of *E-Rater*

E-Rater: What It Is, How It Works

General Approach

"*E-rater*" is the automated scoring system that has been developed to assist in the evaluation of test takers' responses to open-ended essay questions. The current version of the system is capable of assigning scores (on a 0-to-6 scale) to two kinds of essays -- one that requires examinees to discuss an issue and another that requires them to analyze an argument.

E-rater is empirically based. It is "trained" by being fed samples of essays that have been previously scored by human readers. The samples are randomly selected for each essay prompt so as to represent a complete range of possible scores. Essentially, *e-rater* uses natural language processing techniques to duplicate the performance of human readers.

In its attempt to model human readers, *e-rater* uses several subroutines to extract a wide variety of features of the essays that it evaluates. 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. Ordinary (stepwise) least squares linear regression analysis is used to select the set of features that best predicts these scores. A total of 50-60 features are "extractable," but in practice only a subset of the most predictive features, usually about 8-12, are retained and used for each essay prompt. The least squares regression weights for these predictors are applied to each new essay to estimate a score. The score is rounded to the nearest integer from 0 to 6 in order to place it on the scale used by human readers.

E-rater is prompt specific. That is, it is trained for each different essay prompt. Thus, the final set of features used to compute scores for any given prompt is likely to differ somewhat across prompts. Several features are much more likely than others to recur as predictors across different prompts. Some of the same features tend to be predictive of scores for both the issue and the argument prompt types.

Although *e-rater* is empirically based, it is not "blindly empirical." *E-rater* could, of course, take a "brute-force" approach, extracting numerous features indiscriminately. However, not all of the extractable features correspond equally well to the features that human readers are instructed to consider when they score essays. Therefore, *e-rater* features are required not only to be predictive of human readers' scores, but also to have some logical correspondence to the characteristics that human readers are trained to consider. These characteristics are specified in the scoring rubrics that guide human readers when they score essays.

Specific Techniques

Although certain "surface features" (such as the number of words in an essay) are easily extractable, these purely surface characteristics are *not* used by *e-rater*. Instead, the focus is on three general classes of essay features -- structure (or syntax), discourse (organization), and content (or prompt-specific vocabulary).

Structure

Structural characteristics include such notions as "syntactic variety," that is, the use of various structures in the arrangement of phrases, clauses, and sentences. In this category, several specific features have proven to be predictive of human readers' scores. They are:

- the number of subjunctive modal auxiliary verbs (*would*, *could*, *should*, *might*, *may*)

- the prevalence of infinitive clauses ("*To support his expensive hobby*, Phil Atelist would soon need to get a second job at the post office.")

- the proportion of complement clauses ("He felt *that it was well worth the effort."*)

- the incidence of subordinate clauses ("*Although the argument points out that trends are part of our society*, it does not explain why.").

Organization

How well test takers are able to organize their ideas is a major focus for human readers and for *e-rater*. *E-rater* evaluates organization by extracting a variety of rhetorical features (i.e., 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 being developed. (By argument, we mean, generally, 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 and phrases like the following are used to express opinions or beliefs:

- *certainly, clearly, obviously, plainly, possibly, perhaps, potentially, probably, fortunately, generally, maybe, presumably, unless, albeit, luckily, normally, apparently, herein, likely, surely, ideally, undoubtedly, naturally*

- *for certain, for sure, of course, to some extent, above all, if only, in order to, in order for, so that, so as to*

The significance of many of the words listed above is that they signal the start of an "argument" (in the sense that we have described), and the number of arguments developed is a feature that is emphasized by *e-rater*. Arguments sometimes begin simply with *I* or *we*, or with "parallel" words or phrases that signify the start of another line of argument. The latter include the following:

- *firstly, essentially, additionally, first, second, third, another, secondly, thirdly, fourth, next, finally, final, last, lastly, moreover, too, also, likewise, similarly, initially, further, furthermore, first of all, in the first place, for one thing, for a start, second of all, many times, more importantly, most importantly*

The onset of an argument is also signaled by rhetorical words and phrases such as *suppose, supposedly,* and *what if*. The subsequent details are often preceded by words and phrases like:

- *if, specifically, particularly, when, namely, for example, for instance, e.g., in this case, in that case, such that, as well as, in that, such as, about how, in addition, in addition to*

Words or phrases that are used to contrast points of view or to indicate alternative opinions are often found in arguments. These include words and phrases such as:

- *otherwise, conversely, however, nonetheless, though, yet, meanwhile, while, but, instead, although, still, notwithstanding, anyway, unlike, on the contrary, in contrast, by comparison, in any case, at any rate, in spite of, rather than, on the other hand, even then, even if, even though, apart from, instead of*

Some words or phrases signal the presentation of evidence to support an argument. These include *since, because, actually, in fact, after all, as a matter of fact,* and *because of*. Still others signify the stage at which inferences are being made:

- *accordingly, consequently, hence, thus, ultimately, so, thereby, then, therefore, following, after, afterward, afterwards, as a consequence, as a result, if so, if not, as such, according to, in turn, right after*

Yet other words and phrases signal that an argument is being *re*formulated (*alternatively, alternately, that is, in other words, briefly*). In short, *e-rater* looks for linguistic clues (such as logical connections between sentences and clauses) that signify logical or well-ordered thinking.

Content

By means of "topical analyses," *e-rater* also considers the content or vocabulary of the essays it evaluates. These analyses are predicated on the assumption that well-written essays are more responsive (or relevant) to the topic posed than are poorly written essays. Better essays also tend to use more precise and specialized vocabulary. Therefore, the expectation is that, with respect to the words they contain, good 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 number based on the similarity of its content to samples of previously-scored essays. This evaluation proceeds as follows. First, for every essay a determination is made of each word's "contribution" to the essay. (Variations of a word, such as "walks," "walking," or "walked," are considered to be the same word.) A word's relative contribution to an essay is estimated by computing a weight for each word in every essay. These word weights reflect both the frequency of a word in a given essay (relative to the frequency of other words in the essay), and the distribution of the word across other essays. The formula is such that words appearing relatively frequently in an essay will, all other things being equal, receive a higher weight than words appearing less frequently in the same essay. However, words that tend to be used in many essays will get lower weights, all else being equal, than those used in few essays.

In order to assign a numerical value to reflect an essay's content, the word weights for each essay are compared with the weights computed for words in the previously scored training essays. An essay's content feature score is assigned by determining the set of essays that its word weights most closely match. Essays whose word weights correlate most strongly with those computed for highly scored essays will get higher content scores than will essays whose word weights resemble those of weaker essays.

As an example, consider the following (partial) results based on essays analyzing an attempt to convince companies to utilize billboards to increase sales. The two most highly weighted words in essays that received the lowest score (i.e., 1) were "picture" and "percent." Both of these words received much lower weights in essays that got higher scores. On the other hand, the two most heavily weighted words ("recognition" and "local") in the best essays (i.e., 6s) received much lower weights in the poorer essays. The weights assigned to these four specific words in essays at each score level were as follows:

	Score level					
Word	1	2	3	4	5	6
Picture	**.55**	.17	.21	.09	.12	.08
Percent	**.51**	.25	.24	.24	.05	.05
Recognition	.00	.04	.06	.19	.41	**.69**
Local	.05	.11	.11	.20	.16	**.32**

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 by human readers. The formula used to compute the argument content score assigns higher values to essays with many arguments than to those with few.

***E-Rater* Models**

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 (i.e., a set of features that is most predictive of human readers' scores) is specified for each essay prompt. As stated earlier, the set is generally somewhat different for each prompt. Some features are used much more frequently than others in the predictive models. The most frequently occurring are:

(1) content by argument
(2) content by essay
(3) the number of subjunctive auxiliary verbs
(4) the ratio of subjunctive auxiliary verbs to total words in the essay
(5) the total number of argument development terms

The weights assigned to these variables may be either positive or negative. For instance, the *number* of subjunctive auxiliary verbs typically receives a positive weight, while the *ratio* of such words is usually weighted negatively. Although the weights assigned to most of the features discussed earlier are usually positive, some features tend to have negative weights. These include certain words (*because, since, actually*) used to present evidence and the use of pronouns (*I, we*) to begin arguments.

The overview provided here should give the reader some sense of how *e-rater* functions. More detail is available in a number of reports that can be downloaded from the following Web site: http://www.ets.org/research/erater.html. If you are unable to download, contact us and we'll send the papers electronically as Word documents or by regular mail in hard-copy form.

Appendix C

Strategies Employed for Writing Essays Hypothesized to Receive Higher Scores From *E-Rater* Than From Human Readers (and Actual Discrepancies Between Scores)

Prompt type	Hypothesis/strategy	Actual discrepancy
Issue	***The paragraph that is repeated 37 times to form this response probably deserves a 2. It contains the core of a decent argument, possibly the beginning of a second argument, and some relevant references, but the discussion is disjointed and haphazard, with the references and transition words badly misdeployed. E-rater might give too high a rating to the single paragraph because of the high incidence of relevant vocabulary and transition words. However, I've repeated this poor, short, dashed-off response several times to improve the score (as a test-taker could easily do). As far as I can tell, e-rater's score should be linear on the number of copies: it is a linear combination of various features that all seem to be linear on the number of copies. Some features (ratios, cosine measures) remain constant, and the others (counts) are computable by a weighted FSA and hence are linear. Therefore, if I understand correctly, sufficiently many copies should yield an extreme score of 0 or 6, depending on the slope of the linear score function. My guess is that in this example, the slope is positive and a high score will result.***	+ 5.0
Argument	***This paper appears to be a response to the topic, but it merely takes an easily coachable shell, arbitrarily inserts a few key words and phrases from the topic itself (a coachable technique). It then makes a series of paragraphs by simply reordering the same statements in different ways and throwing in some transitional phrases to suggest analytic content & logical organization. I expect e-rater to overrate it.***	+4.0
Argument	***This long and (I hope) fluent essay is one that should get a score of 2 from an experienced reader, since it fails to provide any critical analysis of the argument's logical features and instead, agreeing with the argument, gives an issue-style response. I think that it would be scored higher than a 2 by e-rater, since I have seen various GMAT operational examples in which this very situation has occurred. My faux response is long; it uses many transitional phrases and subjunctive auxiliary verbs; it has fairly complex and varied sentence structure; and it employs content vocabulary closely related to the subject matter of the prompt. All of these features seem to be overvalued by e-rater (even though length per se is not supposed to be a factor).***	+3.5
Issue	***This response is a long one with many transitional phrases and subjunctive auxiliary verbs, and a fair amount of complexity in sentence structure, but it is one that should receive a low score from an experienced reader. It completely misses the point of the prompt and goes off on another tangent; moreover, it is repetitive and somewhat haphazard in its progression from one thought to the next. I think that e-rater would give too high a score because the surface features (listed in the first sentence above) are ones that it values and because the content vocabulary is related to that of the prompt. Moreover, I think that e-rater tends to reward the use of a generic shell such as the one I employed in the first paragraph of the essay*.**	+3.0
Argument	***No argument, just a rambling pastiche of canned text (often repeated for emphasis), possibly relevant content words, and high-falutin function words. Despite the many flaws in diction, I don't think a grammar checker would find much to complain about, if you are using one at all.***	+3.0

Prompt type	Hypothesis/strategy	Actual discrepancy
Issue	***This essay presents a series of pseudo arguments, most of which "fall over" because the writer is dramatically badly informed. In reality this essay should never score more than 2. It has a structure, but the information in the examples is "potty" and the arguments are fallacious. I am hoping that e-rater should pull in a 4 at least because the essay appears adequately structured and the writer does take a position on the issue and has a reasonable control of English in the academic style*.**	+3.0
Issue	***I assume e-rater cannot detect the fact that I just copied the prompt and task for writing using different words, as some ESL writers do in their essays*.**	+2.0
Argument	***My general strategy for the argument essays was to make the essay long and to use more sophisticated language. I tried to write a long and fluent essay that would "seduce" e-rater into giving a high score. Experienced readers should score this essay a 2 because it offers no analysis of the line of reasoning, but instead offers either analysis of tangential issues or the author's own argument on the topic of skater safety. Although the essay contains various syntactic structures, organizational cues, and the "content" is appropriate, the scoring guide does not support a score higher than a "2" because of the lack of analysis of the argument.***	+2.0
Issue	***Thought your computer might like to have a go at this. Perfect grammar and spelling yet semantic garbage.***	+1.5
Argument	***This essay is largely non-sensical to human readers. I believe it will get an inflated score because I based the structure and content of the argument on the benchmark essay for the same topic (the mentioned essay received a score of '6'). The essay is largely a challenge to e-rater's ability to differentiate between efficacious content and nonsense. I went through the well scored essay and edited most every sentence, adding new words usually intended to confound the logic of the arguments without straying from the diction of the topic.***	+1.5
Issue	***I was curious as to e-rater's ability to deem an essay off-topic if the content words were familiar to its database of weighted key-words. So I set out to write an essay that was poorly organized in thought and intention, but tried my best to plug in key structures expected of an issues or persuasive exposition. All the while I maintained usage of key content words I thought would be valued positively by e-rater. In my opinion, if e-rater gives this essay anything but a '0', it has erred.***	+1.5
Argument	***This essay can be described as well-written drivel. All of the points faintly resemble logical analysis, but there is no actual analytic thinking or writing to be found in this essay. I predict that e-rater will love the unremitting use of prompt words, and the almost painfully obvious attempt to include every possible word signaling opinion, start of an argument, conclusion of an argument, etc. My only hope is that the human reader is not fooled. Since the essay is well written and long, and tosses around all the right "buzz" words, this is a real possibility.***	+1.5
Issue	***I attempted to make the core arguments nonsensical by changing key words. It would be interesting to see how the e-rater would cope with this deliberate attempt to create a good essay and change it slightly to make it a really bad essay*.**	+1.5
Argument	***This essay uses a fairly sophisticated syntax and follows a logical progression of thought. However, a human reader will notice the superficiality of the argument, its lack of examples, etc.***	+1.0

Prompt type	Hypothesis/strategy	Actual discrepancy
Issue	***I tried to use catch phrases whenever possible, and most of the information is fabricated. I am counting on e-rater not to know when I am lying and when I am telling the truth*.**	+1.0
Issue	***I tried to use as much of what I knew about the grading scheme as I could. When I wanted e-rater to give a better grade than a human, I tried to use more complex sentence structures, words related to the topic, specific and lower-frequency words, and argument-marking words, and to write at greater overall length. But I also tried to offset this by having the actual claims of the essay be wildly self-contradictory or irrelevant. I should mention that, despite what I believe to be a fairly good command of the language, I am not well-versed enough in grammar to know the differences between the verb forms mentioned in the handout, and so it all broke down into 'complex' and 'simple' for me.***	+1.0
Argument	***Frankly, [a student who wrote like this] should be expelled for plagiarism and stupidity. What I did was to pick a set of meaty looking phrases from the benchmark 5 and 6 essays and string them together using words that you listed under "organization" in "E-Rater: What It Is, How It Works." I have tried wherever possible to make the resulting multiclause sentences as idiotic as possible. I am hoping that the apparent structure and high rate of technical terms should induce e-rater to give the essay at least a 4 if not a 5. Actually it is piffle. It provides absolutely no evidence of understanding and should score 0. I do not suppose that this is a very original approach but it is the sort of thing that a particular kind of end-gaining student might try.***	+1.0
Argument	***I just copied the prompt and task for writing several times using different words as I did in essay one*.**	+1.0
Issue	***Provided additional argument development terms and subordinate clauses*.**	+0.5
Argument	***Provided additional argument development terms and subordinate clauses*.**	+0.5
Issue	***This essay is so convoluted that even I can't follow it. But all the conventions are present and in use. A careful reading elicits a "Huh?" from the reader, since I have continuously flip-flopped between the opinion that choice is everywhere and the opinion that choice is nowhere. It truly makes no sense, and should get a 2 according to the scoring guide. I predict e-rater will give it a 4. Unfortunately, an undiscerning human reader might also give it a 4*.**	+0.5
Issue	***This essay is seemingly addressing the topic, but the arguments have nothing to do with the statement. There aren't a lot of structures used and the vocabulary of the writer is very limited as well. There is no logical connection between the paragraphs and though seemingly there is an introductory and closing sentence of the text, it has not much to do with the body. There is no fluent language used.***	+0.5
Argument	***There is no logical connection between the paragraphs, and the writer is expressing his thoughts about the issue, rather than giving his comments on the argument itself. The vocabulary used in the text is very limited and there are some grammatical mistakes as well, which sometimes make the given sentence incomprehensible. There is not much reflection to what has been mentioned in the argument. Parts of it are only factually stated as if the writer had come up with the statistics.***	0.0

Prompt type	Hypothesis/strategy	Actual discrepancy
Argument	***The structure of this essay is sound, but the matter within the structure is nonsense. The first point is false. The argument has been clearly restricted to the group of those who are injured, and the subgroup who are seriously injured. There is no need to analyze the population. The second point ignored the preamble to the argument that the location of injuries was in parking lots and streets. It missed the point. The third point is just gibberish but has a lot of vocabulary in it. Finally, the conclusion cannot be drawn from the argument in the essay***.	0.0
Argument	***I tried to use as much of what I knew about the grading scheme as I could. When I wanted e-rater to give a better grade than a human, I tried to use more complex sentence structures, frequently use words related to the topic, use specific and lower-frequency words, use argument-marking words, write at greater overall length, but then offset this by having the actual claims of the essay be wildly self-contradictory or irrelevant. I should mention that, despite what I believe to be a fairly good command of the language, I am not well-versed enough in grammar to know the differences between the verb forms mentioned in the handout, and so it all broke down into 'complex' and 'simple' for me***.	0.0
Issue	***This text contains numerous subordinate clauses and modal auxiliary verbs. The structure of the argument is usually revealed for the program by cue words. These high-scoring tactics are applied even when such usage results in sentences than a human reader would find unnatural. However, I have no basis for adjusting the choice of vocabulary to skew the results, because I do not know the pattern of weights assigned to human-graded essays, and the content is graded based on a correlation with those weights***.	−1.5

Note. Several writers gave predictions, but offered no reasons for them.

Appendix D

Strategies Employed for Writing Essays Hypothesized to Receive Lower Scores From *E-Rater* Than From Human Readers (and Actual Discrepancies Between Scores)

Prompt type	Hypothesis/strategy	Actual discrepancy
Issue	***I've used literary metaphors and allusions (Frankenstein, Alfred Hitchcock's concept of "maguffin," Catch-22), and I've avoided all obvious content words. The connections and development are subtle and implicit, and there are no transitional giveaways or phrases that announce that an argument is about to happen. If e-rater is not sophisticated enough to recognize this, the essay might get a 3. If e-rater gives this a high score, I might become a true believer.***	−2.0
Issue	***I wanted to see how e-rater would treat the use of an extended metaphor as a means of development, since the content vocabulary might differ from that encountered in most of the samples used to "train" it. I deliberately kept the response short; I also tried to avoid the use of very obvious structural signposts ("first," "second," "in conclusion," etc.).***	−2.0
Argument	***This response probably deserves a 6 or at least a 5. However, it has a number of features that may give it a lower score: few classical transition words, complement clauses, etc., some quoting from the problem, as may appear in some of the more desperate low-scoring training responses, many modal auxiliaries, as well as "maybe" and "actually," few typos that probably match against misspellings in a few of the low-scoring training responses (note that misspelled words have high IDF weight). Comparatively low cosine measure with other responses of any score. (The largest coordinates are probably for unusual words drawn from the analogy, like "sultan.") Thus, the cosine measure has fewer words (either positive or negative) to influence its decision. While this does not introduce a bias toward matching the low-scoring responses, it may reduce the reliability of matching with the high-scoring responses. That is, I imagine ArgContent's output will be a bit more random than for an essay that falls squarely into the training class; and note that if ArgContent were fully random, its average score would be 3.***	−1.0
Issue	***This text has a simplified sentence structure with very few subordinate clauses, which has the unfortunate side-effect of causing the tone to appear dogmatic to a human ear. The discourse structure is seldom revealed by cue words. Instead, the human reader would deduce the logical connection between sentences based on meaning. Conjunctions such as "but" or "and" are used, because they are likely to score lower than "although", "furthermore", etc.***	−1.0
Issue	***The essay is focusing on a few important items and is giving a list of possible topics that could be discussed under the issue. It presents a large vocabulary, and makes only few grammar mistakes. The paper has a beginning, a body and a summary, which includes a conclusion, but that is not a direct reflection to the quote itself. He has developed a position on the issue and gives support to his argument by mentioning examples. It uses the language fluently, but some of the sentences seem to be driven away from the topic. There are several sentence structures used in the paper.***	−1.0

Prompt type	Hypothesis/strategy	Actual discrepancy
Issue	***This brief, dense, seven-sentence response deftly makes several points and presumably deserves a 6. The formal structure unfolds simply enough that it gets away with very few transition words -- which I suspect is good for the reader, bad for e-rater. Specifically, almost every sentence lays out a new argument that builds on the previous argument, alluding back to the previous argument's "content" rather than using a transition word. This serial development is interrupted only by two instances of counter argument (signaled by "If you object"; "It is true ... but") and one instance of parenthesis. Moreover, the response is stronger because it recognizes a number of subtleties in passing, without interrupting the flow to couch them as separate arguments. (See, e.g., the "competitive success" sentence, which alludes to the case of unfair competition, distinguishes nonconformity from personal originality, and unifies various pursuits as competition over "market share or mind share" while distinguishing them from other kinds of competition, such as baseball.) I should mention that this is essentially a natural and favored response on my part -- i.e., something like this might show up on a real exam, although I personally would not have been able to achieve this degree of compression in 15 minutes. I wrote it freely with the intention of tweaking it somehow to lower the score, but concluded that e-rater might have problems anyway and so tweaked it very little.***	−0.5
Argument	***This response is [relatively short], but it is clearly substantive and does provide accurate analysis of the argument's logical features. I believe that e-rater would score it lower than a human reader would, because it is brief, it uses only one paragraph, it has few transitional phrases or subjunctive auxiliary verbs, and in general it lacks the organizational features that e-rater seems to reward.***	−0.5
Argument	***This paper recognizes very clearly all the major flaws with the argument and exposes them by pretending to agree with them (a Jonathan Swift approach, combining parody and argument ad absurdum). I'd give such a paper a 5 or 6. Since it analyzes the reasoning by considering a parallel example, it has very few words and constructions in common with the topic or, I would think, with other papers on the topic. A decent reader should see the relevance of the discussion immediately (and perhaps even enjoy the novel approach).***	−0.5
Argument	***This essay presents a well-written, precise analysis, but I predict that e-rater will not be able to recognize that it's a 6 (or high 5) essay. I used virtually none of the words from the prompt (very tough to do!), and provided almost no words or phrases that conventionally signal opinion, the start of an argument, subsequent details of an argument, contrasting points of view, connections, etc. In other words, the plot thickens, but there are no signal phrases in this essay to indicate movement. There are no crutches or giveaways here; the reader has to read every word and find the very implicit, very organic connections. Nonetheless, the essay is a perceptive critique of the argument (I know; I've read hundreds of essays on the topic), and this essay fits the scoring guide description of a 6.***	0.0
Issue and Argument	***I attempted to make the essays more elliptical, limiting words and personalizing the texts***.	0.0 for Issue 0.0 for Argument

Prompt type	Hypothesis/strategy	Actual discrepancy
Issue	***I organized the essay around a "shell" of the sort that coaching schools might teach. The shell here is: The topic claims that no one can succeed by conforming to conventional practices or ways of thinking. This is a very complex issue, one that must be examined from all sides ("from one perspective," "from another perspective," "on the one hand," "on the other hand," "in the final analysis"). The shell, although somewhat forced, is used effectively and logically. I've deliberately kept the syntax less varied [than in the other essay I wrote], and I've used fewer transitions. Nevertheless, the reasons and examples are relevant and persuasive, and clearly related to the points under discussion, unlike in [my other] essay. I predict that e-rater will undervalue this essay, scoring it in the lower half, whereas I think experienced readers should score this essay a 5***.	0.0
Argument	***When I wanted e-rater to give a worse grade than a human, I kept both my sentences and the essays quite short and simple. I avoided using argument markers and words which were either major content words from the prompt or their synonyms (which led to some rather awkward circumlocutions, somewhat in conflict with my desire for a simple sentence structure). I tried to counteract these deficiencies by sometimes having a single thought reminiscent of the complex thinking behind a complicated sentence appear in two adjacent sentences. I also tried to structure the argument in a clear and obvious manner, bringing up important points in the process. I should mention that, despite what I believe to be a fairly good command of the language, I am not well-versed enough in grammar to know the differences between the verb forms mentioned in the handout, and so it all broke down into 'complex' and 'simple' for me***.	0.0
Argument	***I deleted [from another essay] (a) the words indicating the beginning of each argument (first of all ...) (b) 'apparently' in paragraph two, 'specifically' in paragraph four, and 'initially' in paragraph five (as being redundant), but I assume e-rater is designed to give weight to these words***.	0.0
Issue	***This essay uses a threaded/parallel development which I think the computer will score lower because its development does not follow standard argumentative lines. It might "confuse" e-rater that the paragraphs do not mix linearly. Too, it uses a lengthy quote from Shakespeare -- something I think that a human reader would far more appreciate than e-rater***.	0.0
Issue	***This essay was designed to evoke an e-rater grade lower than it deserves. I avoided using the catch phrases identified by e-rater as good organization markers and I tried to use company names that would cause e-rater to have punctuation problems.***	+0.5
Argument	***[I wrote this] in memoriam to every Welsh bar bore I've ever met. The content is obscured by the writer's curious rhetorical style, but somewhere beneath it all, under a few thick layers of local sentiment and bred in the bone socialism, he has a point. That is, the argument is flawed because it only considers the equipment and not the people who use it. Provided you had consumed sufficient Welsh bitter, you might find this guy convincing. I would score him a 3 because he identifies one important feature, develops his ideas to an extent employing his own particular logic and has a reasonable grasp of English even if his is incapable of writing in an academic manner. I am hoping that e-rater will think he is completely off topic and give him 0. Failing this, he may score 1 because I suspect that e-rater will be unable to follow an anecdotal argument***.	+0.5

Prompt type	Hypothesis/strategy	Actual discrepancy
Argument	***The paper is well-written and thoughtfully discusses the topic. It reflects on two important parts of the argument and gives a relatively detailed opinion about it. The thoughts are logically organized, even follow the sequence of the arguments in the quotation, and connect the ideas clearly. There are a few flaws in the text, and the vocabulary used is not very large.***	+0.5
Issue	***When I wanted e-rater to give a worse grade than a human, I kept both my sentences and the essays quite short and simple and avoided using words which were either major content words from the prompt or their synonyms (which led to some rather awkward circumlocutions, somewhat in conflict with my desire for a simple sentence structure). I also avoided use of argument markers, but counteracted these deficiencies by sometimes having a single thought reminiscent of the complex thinking behind a complicated sentence appear in two adjacent sentences, and structured the argument in a clear and obvious manner, bringing up important points in the process. I should mention that, despite what I believe to be a fairly good command of the language, I am not well-versed enough in grammar to know the differences between the verb forms mentioned in the handout, and so it all broke down into 'complex' and 'simple' for me.***	+1.0
Argument	***My general strategy was to [write a relatively short essay] and use relatively unsophisticated language. I tried to write an essay that would be undervalued by e-rater -- I predict that e-rater will score this essay significantly lower than experienced readers. This essay differs from any I've seen on the topic in that it begins with a linguistic analysis of the meaning of "need for more protective gear." The essay lacks the sophisticated diction and syntactic variety that e-rater seems to prize. and it contains fewer organizational cue words than that contained in typical, upper-third papers. Nevertheless, the organization is clear and logical, and the analysis is quite thorough. I predict that experienced readers would give this an upper-third score, which is what the scoring guide requires.***	+1.5
Issue	***To put it in transatlantic terms, this is an intelligent arts major who has been bullied into doing a social sciences course against [his] will and is making a considered and creative attempt to fail so badly so no one will ever try to make him do so again. Possibly he has also heard that e-rater is being used and is offended by it. Hence the essay is written throughout in "Arabian Nights" style. I hope this will make e-rater see the essay as off topic and score it 0. However, the Grand Visar has several excellent arguments. To translate, these are: We hear a lot about unconventional successes because they make good copy. On the other hand, more conventional people often produce successes which last because they are sustainable, conventions often arise for good reasons, and the secret of success is to choose your conventions carefully. As examples, the Grand Visar presents good manners and e-rater itself, which is clearly designed to reward conventional essays. I would score this essay 5 (or possibly 4 depending on how rigid you are in your definition of "appropriate vocabulary").***	+2.5
Issue	***I just deleted [from another essay] (a) the words indicating the beginning of each argument (firstly ...) and (b) 'for instance.' As a rater, I know that every new paragraph could be an argument, but I doubt if e-rater does the same; and I consider 'for instance' redundant in the fourth paragraph, but e-rater might give a point for it.***	+3.0

Note. Several writers gave predictions, but offered no reasons for them.

54020-007952 • Y31P.4 • Printed in U.S.A.
I.N.991141