Toward Theoretically Meaningful Automated Essay Scoring

Randy Elliot Bennett

Educational Testing Service

Princeton, NJ

and

Anat Ben-Simon

National Institute for Testing and Evaluation

Jerusalem, Israel

December 2005

This report was produced with funding from the National Center for Education Statistics, Institute for Education Sciences, US Department of Education, grant #R902B04006.

Acknowledgments

This study could not have been conducted without the assistance of many individuals to whom we would like to express our most sincere gratitude. Ten experts from the writing community played a central role in suggesting weights for the e-rater features, as well as in educating us about ways for improving the NAEP rubrics, e-rater, and automated scoring more generally. Yigal Attali provided consultation on procedures for scaling, establishing cut points, and weighting e-rater scores. Jill Burstein gave advice on various alternative approaches to training e-rater. Chi Lu created the scoring models for e-rater-E and e-rater-H, and produced the scores. Fred Yan and Bruce Kaplan created the data sets for analysis. Mark Shuvman performed the many, extensive data analyses needed for the project. Several individuals, including Henry Braun, Shelby Haberman, Neil Dorans, and Charlie Lewis offered advice on statistical procedures and data analysis. Finally, Yigal Attali, Dan Eignor, and Don Powers gave helpful comments on the final report and Yehudit Meroz designed and executed the layout.

Any opinions expressed are those of the authors and not necessarily of Educational Testing Service or NITE.

Executive Summary

Automated essay scoring has the potential to reduce processing costs, speed up the reporting of results, and improve the consistency of grading. This study evaluated a "theoretically driven" method for scoring NAEP writing assessments automatically. The method would be usable for any future NAEP writing assessment conducted on computer or conducted with emerging technologies that allow handwriting to be digitized and translated to type. Such a method, if successful, could produce a means for NAEP to score essay responses automatically in a way that can be linked explicitly to the characteristics of good writing.

Existing commercial programs for automated essay scoring have generally used writing features that are empirically weighted to predict the scores of human raters. The selected writing features may or may not have any direct connection to writing theory. This study used variations of an existing commercial program, e-rater®, to compare the performance of three approaches to automated essay scoring: a *brute-empirical* approach in which variables are selected and weighted solely according to statistical criteria, a *hybrid* approach in which a fixed set of variables more closely tied to the characteristics of good writing was used but the weights were still statistically determined, and a *theoretically driven* approach in which a fixed set of variables was weighted according to the judgments of writing experts.

The research questions concerned (1) the reproducibility of weights across writing experts, (2) the comparison of scores generated by the three automated approaches, and (3) the extent to which models developed for scoring one NAEP prompt generalize to other NAEP prompts of the same genre. Data came from the NAEP Writing Online study (Horkay, Bennett, Allen, & Kaplan, 2005), which included the responses of 1,255 8th grade students to two essays, and from the main NAEP 2002 writing assessment, from which 300 responses to each of four essays were employed. Weights were provided by two committees of writing experts.

Results showed that experts initially assigned weights to writing dimensions that were notably more similar across the two committees than to the empirically derived weights used by the hybrid approach. When one committee was shown the empirical weights and the other committee was not, the differences between the committees increased, with the committee shown the weights moving closer in its judgments to the weights of the hybrid approach. As a consequence, each committee's weights was used separately in the analysis.

The various automated approaches were compared with respect to their relations with human scores, their relations with other indicators, their functioning in NAEP reporting groups, and the resolution of large machine-human score discrepancies. The theoretical approach based on committee judgments informed by the hybrid's empirical weights generally did not operate in a markedly different way from the brute empirical or hybrid approaches. In contrast, many consistent differences with those approaches were observed for the theoretical approach based on the judgments of the committee that was *not* informed of the empirical weights. For example, this theoretical approach produced mean scores that were significantly lower than human scores; correlated less with human scores than did the hybrid version; had considerably lower exact agreement with humans than did either the brute empirical or hybrid versions; and had a lower between-prompt correlation than observed for human scores.

With respect to generalizability to other prompts, the theoretical approach based on committee judgments informed by empirical weights fared less favorably than the brute empirical and hybrid approaches, but usually by small amounts. In contrast, the theoretical approach based on

the judgments of the committee *not* informed by the empirical weights showed more and larger differences.

Should NAEP decide to use automated scoring in future online writing assessments, empirical weights might provide a useful starting point for expert committees, with the understanding that the weights be moderated only somewhat to bring them more into line with theoretical considerations. Under such circumstances, the results may turn out to be reasonable, though not necessarily as highly related to human ratings as statistically optimal approaches would produce.

Introduction

NAEP spends significant time and monetary resources for scoring essay responses. At some point in the not-too-distant future, NAEP writing assessments will be delivered on computer or possibly through electronic pen or notebook technologies that allow handwritten responses to be digitized for machine processing. If NAEP essays could be scored automatically, results might be reported sooner, money saved, and grading consistency improved.

At least four commercially available programs for automated essay scoring exist. In principle, these programs may be less susceptible to the systematic biases and random errors that human raters make (e.g., fatigue, halo, handwriting, and length effects and the effects of specific content). The research on automated essay scoring suggests that these programs produce grades that compare reasonably well with the scoring judgments of human experts (Keith, 2003).

Although automated scoring programs function reasonably well, the methods they use to arrive at scores are, from the perspective of many in the writing and measurement communities, conceptually weak (Bennett, 2006; Cheville, 2004). This weakness is most apparent in two ways. First, the specific features of student writing used to generate scores are usually not linked to good writing in any finely articulated, theoretically grounded way. Second, writing features are typically combined to form scores solely by statistical techniques, most often a multiple regression of human scores from a training sample of essays onto computed essay features. Because this regression is usually estimated for each writing prompt separately, not only may the feature weights differ from one prompt to the next but the features themselves may vary. The result is a selection and weighting of features that, while optimal for predicting the scores of a particular group of human readers, may make little sense to writing educators more generally.

Two fundamental questions underlie the current study. First, if a computer can produce scores for essay responses that are comparable to human scores, do we care how the machine does it? Second, can we capitalize on the fact that a computer can simultaneously process many writing features by selecting and combining those features in a more theoretically defensible way? This study is motivated by the belief that the answer to both questions is "yes." We need to care how the machine computes its scores because, if automated scoring is done in a substantively and technically defensible way, it should:

1. bolster construct validity by making explicit the links between the features of student responses and the scores those responses receive,
2. allow for more meaningful and detailed descriptions of how groups differ in their writing performance, and
3. make results more credible to writing educators, parents, and policy makers.

Literature Review

Automated Essay Scoring

Automated essay scoring (AES) is "the ability of computer technology to evaluate and score written prose" (Shermis & Burstein, 2003, p. xiii). AES may offer important advantages over conventional grading, including greater objectivity (i.e., a specifiable algorithm for scoring), standardization (i.e., the same criteria applied to all responses), and efficiency (i.e., quick and inexpensive production scoring). The potential value of this technology has been recognized by

the National Commission on Writing in America's Schools and Colleges, which recommends research and development of AES systems for standardized tests (National Commission on Writing, 2003, pp. 30-31).

Most AES systems attempt to mimic, as closely as possible, the scores produced by human raters. This outcome is achieved in the following way. First, human readers grade a training sample of up to several hundred responses. Next, an AES program produces a scoring model by identifying a set of features and weights that best predicts the human ratings in the training sample. This scoring model is then cross-validated in a second sample of human-scored essays. Once the scoring model is functioning satisfactorily, new responses can be automatically scored by extracting the relevant features and applying the weights.

Though most AES systems use the same general training process, their particular approaches to scoring vary in fundamental ways. Key to the current study are three specific aspects of scoring: (1) the type of lower-level features used by the system and, in particular, their relationship to writing characteristics grounded in a theoretical model; (2) the grouping of these features into higher-level writing dimensions; and (3) the procedure by which these features are weighted in the scoring model to produce scores.

AES systems can be roughly classified into two categories: systems based predominantly on brute-empirical methods and systems based on hybrid methods. AES systems based on brute-empirical methods typically extract a large variety of linguistic features from an essay response. These features will often have no direct, intuitive link to writing theory. In addition, both the features used in the final scoring model and their weights will be empirically derived. Finally, the features may be collapsed to produce a smaller number of dimension scores but the assignment of features to dimensions may be more a matter of convenience than of theoretical principle.

In contrast, systems based on hybrid methods typically use a smaller set of features more closely related to a theoretically derived conception of the characteristics of good writing. This theoretical conception may also drive the assignment of features to higher-level dimensions. But similar to the brute-empirical approach, the features are usually weighted empirically to best predict human scores.

The following is a brief description of the four leading commercial essay-scoring systems--PEG (Project Essay Grade), IntelliMetric, the Intelligent Essay Assessor, and e-rater®--in terms of these two categories.

PEG was the first computer program developed for essay scoring. Ellis Page created the original version in 1966 (Page, 1966). This version used approximately 30 features (called "proxes") that served as stand-ins or proxies for intrinsic writing qualities (called "trins"). Most features were quantifiable surface variables such as average sentence length, number of paragraphs, and counts of other textual units. The statistical procedure used to produce feature weights was a simple multiple regression.

A revised version of the program was released in the 1990s. This version uses such natural language processing tools as grammar checkers and part-of-speech taggers (Page, 1994, 2003; Page & Petersen, 1995). As a result, this version appears to extract richer and more complex writing features said to be more closely related to underlying trins. A typical scoring model uses 30-40 features. In a recent study, PEG provided, in addition to a total essay score, dimension

scores for content, organization, style, mechanics, and creativity. This innovation was introduced to provide more detailed feedback about students' strengths and weaknesses (Shermis, Koch, Page, Keith, & Harrington, 2002). Exactly what features are used to compose PEG's dimension and total scores is not, however, divulged. As a result, it is difficult to determine whether the current version of PEG is more an example of the brute-empirical or hybrid approaches to automated scoring.

IntelliMetric Engineer (1997) was developed by Vantage Technologies for the purpose of scoring essays and other types of open-ended responses. IntelliMetric is said to be grounded in a "brain-based" or "mind-based" model of information processing and understanding (Elliot & Mikulas, 2004). This grounding appears to draw more on artificial-intelligence, neural-net, and computational-linguistic traditions than on theoretical models of writing.

For any given essay prompt, IntelliMetric uses a training set to extract some 400 features from student responses, identify an optimal set of predictors, and estimate weights to produce a scoring model (Elliot & Mikulas, 2004). The 400 features fall into discourse/rheotorical, content/concept, syntactic/structural, and mechanics classes, though the specific nature of the features in each class is not publicly disclosed.

Five dimension scores are reported:

1. Focus and unity: indicating cohesiveness and consistency in purpose and main idea
2. Development and elaboration: indicating breadth of content and support for concepts advanced
3. Organization and structure: indicating logic of discourse, including transitional fluidity and relationship among parts of the response
4. Sentence structure: indicating sentence complexity and variety
5. Mechanics and conventions: indicating conformance to English language rules

The mapping of feature classes to score dimensions is such that all feature classes contribute to all score dimensions (Elliot, 2003, p. 73), a patently atheoretical formulation. Along with the weighting of features to maximize the prediction of human scores, this mapping seems to put IntelliMetric squarely into the brute-empirical category.

The Intelligent Essay Assessor (IEA) (1997) was created by the University of Colorado (Landauer, Foltz, & Laham, 1998). In contrast to other AES systems, IEA's approach focuses primarily on the evaluation of content. The approach is accompanied by a well-articulated theory of knowledge acquisition and representation (Landauer & Dumais, 1997) and is heavily dependent on Latent Semantic Analysis, a mathematical method that comes from the field of information retrieval (Foltz, 1996; Landauer, Laham, Rehder, & Schreiner, 1997; Landauer et al.,1998). The underlying assumption of the method is that a latent semantic structure (semantic space) for a given set of documents or texts can be captured by a representative matrix that denotes the core meaning or content of these texts. The method is based on a factor-analytic model of word co-occurrences. In this method, information generated from a variety of content-relevant texts (e.g., subject-matter books) is condensed and represented in a matrix that defines a "semantic space" capable of explicitly relating words and documents. The word-document association in this matrix is represented by a numerical value (weight) that is conceptually similar to variable loadings on a set of factors in factor analysis. In the context of essay scoring,

the specific content of an essay is important to the extent that it matches, in the semantic space, other essays of a given score level.

IEA usually provides scores for three dimensions, in addition to a total score:

1. Content: assessed by two features generated from Latent Semantic Analysis, quality and domain relevance
2. Style: assessed by features related to coherence and grammaticality
3. Mechanics: assessed through punctuation and spelling features

IEA's total score is computed from a hierarchical regression of human scores onto the dimension scores.

Although created for the assessment of content knowledge, IEA is also used to evaluate writing skill. In this context, IEA's approach seems to qualify as a hybrid because its analysis of content is theoretically grounded, and content is a key factor in most scoring guides.

e-rater (1997) was developed by Educational Testing Service (Burstein et al., 1998). Version 1 computes approximately 60 linguistically based feature scores from which a subset is selected through step-wise regression. This subset usually includes only 8-12 features for any given prompt. The heavy dependence on relatively low-level linguistic features (e.g., the number of auxiliary subjunctives) and on step-wise regression suggests that this version of e-rater represents a brute-empirical approach very well.

In 2003, a new version (version 2) was created (Attali & Burstein, 2005; Burstein, Chodorow, & Leacock, 2004). This version uses a fixed set of 12 features, many of which are not represented in the first version, that are more intuitively related to the characteristics of good writing. These features can be grouped into five dimensions which, although not used in scoring, are helpful in understanding what the program's developers intend it to measure. The five dimensions, described in Table 1, are Grammar, usage, mechanics, and style; Organization and development; Topical analysis (i.e., prompt-specific vocabulary); Word complexity; and Essay length (Attali & Burstein, 2005). In operational use to date, weights have usually been derived empirically. The primary exceptions to this generalization are for substantively counterintuitive weights, which may be set to zero, and for essay length, which is fixed judgmentally so not to overemphasize the influence of this feature on score computation. The coupling of a more theoretically motivated feature set with the empirical derivation of weights makes for a hybrid approach to scoring.

Validity Issues in AES

Yang, Buckendahl, Juszkiewicz, and Bhola (2002) classify validation approaches for automated scoring into three categories: (1) approaches focusing on the relationship among scores generated by different scorers (human and computer), (2) approaches focusing on the relationship between test scores and external measures of writing, and (3) approaches focusing on the scoring process.

The relationship between human scores and computer-generated scores has been examined for all four AES systems. Consistent with their design to optimize the prediction of human scores, relatively high agreement between the computer and human readers has generally been reported. See Table 2 for representative results.

TABLE 1

Writing Dimensions and Features in e-rater v2

Dimension	**Feature**
Grammar, usage, mechanics, & style	1. Ratio of grammar errors to the total number of words
	2. Ratio of mechanics errors to the total number of words
	3. Ratio of usage errors to the total number of words
	4. Ratio of style errors (repetitious words, passive sentences, very long sentences, very short sentences) to the total number of words
Organization & development	5. The number of "discourse" units detected in the essay (i.e., background, thesis, main ideas, supporting ideas, conclusion)
	6. The average length of each element as a proportion of total number of words in the essay
Topical analysis	7. Similarity of the essay's content to other previously scored essays in the top score category
	8. The score category containing essays whose words are most similar to the target essay
Word complexity	9. Word repetition (ratio of different content words to total number of words)
	10. Vocabulary difficulty (based on word frequency)
	11. Average word length
Essay length	12. Total number of words

Note. Derived from Attali and Burstein (2005).

TABLE 2

Selected Studies Comparing Interrater Reliability to Computer-Rater Reliability

System	Author	Test	Sample Size	Human-Human *r*	Human-Computer *r*
PEG	Page & Petersen, 1995	*Praxis*™ (72 prompts)	300	.65 (average *r* among each pair of 6 ratings)	.74 (average *r* of computer with each of 6 ratings)
PEG	Petersen, 1997	GRE® (36 prompts)	497	.75	.74-.75 (1 rater)
PEG	Shermis, Mzumara, Olson, & Harrington, 2001	English placement test (20 prompts)	617	.62 (median *r* among each pair of 6 ratings)	.71 (*r* with the average of 6 ratings)
PEG	Shermis, Koch, Page, Keith, & Harrington, 2002	English placement test (1 prompt)	386	.71 (median *r* among each pair of 6 ratings)	.83 (*r* with the average of 6 ratings)
Intelli-Metric	Elliot, 2001	K-12 norm-referenced test	102	.84	.82 (average *r* of computer with each of 2 ratings) .85 (*r* with the average of 2 ratings)
IEA	Landauer, Laham, & Foltz, 2003	GMAT® (1 prompt)	292	.86	.84 (1 rater)
		GMAT (1 prompt)	285	.88	.85 (1 rater)
IEA	Landauer, Laham, Rehder, & Schreiner, 1997	GMAT	188	.83	.80
IEA	Foltz, Laham, & Landauer, 1999	GMAT	1,363	.86-.87	.86
e-rater	Burstein et al., 1998	GMAT (13 prompts)	500-1,000 per prompt	.82-.89	.79-.87 (1 rater)
e-rater	Burstein & Chodorow, 1999	TWE® (2 prompts)	270	.69	.75

Note. Praxis is a teacher licensure test. GRE = Graduate Record Examinations®. GMAT = Graduate Management Admission Test®. TWE® = Test of Written English™. The number of prompts and human raters is given where available.

Though high computer-rater agreement is a desirable and perhaps necessary feature of any AES system, it is not a sufficient quality criterion (Bennett, 2006; Cizek & Page, 2003). Unfortunately, studies employing external criteria--Yang et al.'s (2002) second category--are less common. The available studies have used one or more of the following criteria: multiple-choice tests, grades in courses dependent on writing, teachers' ratings of students writing skill, self-evaluations of students' writing skill, and expert-rated essays. Most of these analyses have yielded encouraging, if sometimes incomplete, results because of the limited nature of the external criteria used in any given case (e.g., Elliot, 2001; Landauer et al., 2001; Petersen, 1997; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002).

Validation studies focusing on Yang et al.'s (2002) third category, scoring process, are rare indeed. Since all commercial AES systems use some degree of data-driven statistical procedure to generate their scoring models, additional empirical and theoretical examinations are needed to establish the meaningfulness of these models. Yang et al. emphasize the importance of using descriptive and qualitative approaches to evaluate the automated scoring process. Such approaches can involve analysis of the patterns and nature of disagreement between computer and expert ratings, or identification of differences between human and computer scoring models with regard to writing features and their weighting. More specifically, writing experts can, and arguably should, be used to:

1. judge the relevance of the computer-generated features to the target construct,
2. identify extraneous features, as well as missing ones, and
3. evaluate the appropriateness of the weights assigned to the features.

AES and Writing Theory

Despite the fact that all four commercially available AES programs are being used to assess writing skill, their scoring approaches have limited grounding in writing theory. Though some of the approaches link computer-generated features to characteristics of good writing (see Table 3 for commonly cited characteristics), these approaches typically do not explicitly link specific features to the writing attributes embedded in the rubrics for a particular testing program. This absence is in part due to the fact that developers intend their automated scoring systems to be general enough for a wide variety of writing assessments. In operational practice to date, the linkage to any given assessment has been achieved empirically by the regression of training scores onto computed features. To maximize agreement with human scores, these systems most often use a separate training sample--and, thus, produce a unique scoring model--*for each writing prompt*. Even though the models may vary simply because of differences in the samples of readers or examinees used for training with a particular prompt, writing experts may never be asked to inspect the data-driven features or weights to ensure their substantive appropriateness. As a result, the definition of what makes for good writing may vary from one prompt to the next and the examinee that responds consistently across prompts by incorporating the same features to the same degrees may not receive the same score on each response. This outcome would not seem to be the intended result: In most large-scale assessments, a single rubric is used for scoring all prompts within a genre (though minor adaptations of a rubric may be made to explicate how it should be applied to each prompt).

TABLE 3

Commonly Cited Characteristics of Good Writing

Content	**Rhetorical Structure/ Organization**	**Style**	**Vocabulary**	**Syntax & Grammar/ Mechanics**
Relevance	Paragraphing	Clarity	Richness	Sentence complexity
Richness of ideas	Coherence	Fluency	Register	Syntactical accuracy
Originality	Cohesion		Accuracy	Grammatical accuracy
Quality of argumentation	Focus		Appropriateness to written language	Spelling

Note. Derived from Connor (1990) and Johnson, Penny, and Gordon (1999).

Study Objective

The objective of this study is to lay the groundwork for a more theoretically driven approach to AES that, inYang et al.'s (2002) terms, is concerned with scoring process as well as with the empirical relations of scores. The practical importance of this theoretically driven approach is in potentially providing a more credible and educationally meaningful method for automatically scoring writing assessments, which NAEP can apply once it begins collecting essay responses in digital form.

In line with this objective, the study addressed three research questions:

1. To what extent are judgmentally determined weights reproducible? Some degree of reproducibility across experts and expert committees is required if the underlying basis for scoring is to have conventional meaning.
2. How do the approaches to automated scoring compare to one another in their relations to human scores and to other indicators? A theoretically driven approach should not be expected to relate to human scores as highly as a statistically optimal approach to human score prediction. Any loss in empirical validity, however, will need to be practically small if the use of the theoretically driven approach is to be preferred on a substantive basis.
3. How well does the theoretically driven scoring model developed for one NAEP prompt generalize to other NAEP prompts of the same genre? Some significant degree of generalizability across prompts in a genre should be expected if the judgmentally generated feature weights have broader theoretical meaning.

Method

Participants

The primary data set came from the NAEP Writing Online (WOL) study (Horkay, Bennett, Allen, & Kaplan, 2005). WOL study data were collected in spring 2002 from 1,255 eighth-grade students taking a writing test on computer. In the current study, these data were used in the creation of scoring models and to compare the various automated scoring approaches (research question 2 above).

In addition to the WOL data, a secondary source of data was the eighth-grade main NAEP 2002 writing assessment. From this latter data set, approximately 300 responses to each of four prompts were randomly drawn and key-entered (where each prompt was responded to by a different sample of students). These data were used to test the generalizability of the theoretically driven models created for automatically scoring the two WOL prompts (research question 3 above).

Instruments

As part of the WOL study, the 1,255 students in the primary data set had taken an online writing test consisting of two essay prompts, one informative and one persuasive (see Appendix A for the prompts). Background and demographic information was also collected for each student from questionnaires and school sources.

The data from the online writing test included the raw text responses, one human score for each response, and a second human score for a random sample of the responses. The data file also contained main NAEP 2002 writing plausible values for one nationally representative subset of the sample (N = 687 students) and main NAEP 2002 reading plausible values for the other, nonoverlapping, nationally representative subset (N = 568 students). The writing plausible values were *not* based on the prompts administered in WOL, but rather on a different pair of prompts the students responded to as part of the main NAEP 2002 writing assessment.[1]

The secondary data set from the 2002 main NAEP writing assessment included the raw text responses, at least one human score for each response, and a second human score for a random sample of the responses.

The two examinee data sets are summarized in Table 4.

Automated Essay Scoring Approaches

Three automated scoring approaches were implemented using e-rater versions 1.3 and 2.1 (v1.3 and v2.1). e-rater v1.3 was used to represent a brute-empirical approach and is denoted throughout the remainder of this report as "e-rater-E." Two configurations of e-rater v2.1 were used to represent the hybrid and theoretically driven approaches and are denoted as "e-rater-H" and "e-rater-T," respectively. Table 5 summarizes the three approaches.

[1] To address its research questions, the WOL study intentionally used two nationally representative samples, one drawn from the main NAEP reading assessment and one from the main NAEP writing assessment. Also, the study sample drawn from the main NAEP writing assessment was purposefully limited to students taking a pair of prompts different from the ones administered in WOL. See Horkay et al. (2005) for details.

TABLE 4

Examinee Data Sets Used in the Study

Data Set	Sample	Main Data Elements
NAEP WOL study	1,255 8th grade students	Responses to two essay prompts One or two human scores for each essay response Demographic and background information Main NAEP writing plausible values (N = 687) Main NAEP reading plausible values (N = 568)
Main NAEP 2002 writing assessment	4 samples of 300 8th grade students each	Responses to four essay prompts (one prompt taken by each sample of n = 300) One or two human scores for each essay response

TABLE 5

Three Scoring Approaches as Operationalized by Two Different Versions of e-rater

Scoring	Designation	Description
Brute-empirical	e-rater-E	Operationalized through e-rater v1.3. Computes approximately 60 linguistically derived feature scores for each essay response. Uses step-wise regression to select a subset of features and feature weights that optimally predict human holistic scores in a training set. Typically produces a unique scoring model for each essay prompt.
Hybrid	e-rater-H	Operationalized through e-rater v2.1. Computes a fixed set of 12 features designed to capture five dimensions theoretically related to good writing. Uses hierarchical regression to weight all features (except essay length) to optimally predict human holistic scores in a training set. Typically produces a unique scoring model for each essay prompt (though a single model for multiple prompts can also be empirically derived).
Theoretically driven	e-rater-T	Operationalized through e-rater v2.1. Computes a fixed set of 12 features designed to capture five dimensions theoretically related to good writing. Uses a committee of writing experts to select from among the 12 features and determine weights for the chosen features. In principle, can produce a single scoring model for all prompts within a genre because a single set of features and weights can be designated for each genre.

As should be clear from the table, the brute-empirical approach uses features and weights that were chosen primarily for computational linguistic and statistical reasons, with no clear link to writing theory. The hybrid approach improves on this method by including features that can be better linked theoretically to good writing (see Appendix B for a description of the features). However, because in practical applications to date this approach has generally weighted features statistically, the importance of particular features may be different than theory would suggest. Finally, the theoretically driven approach allows weights and, to a lesser extent, features to be determined through the judgments of writing experts.

Procedure

The study procedure involved four stages, each of which informed one or more research questions. In the first stage, dimension and feature weights were generated by two expert committees. These dimension and feature weights were used in addressing all three research questions. In the second stage, the automated approaches were applied to the WOL data to answer the question of how the automated approaches compared to one another. In the third stage, experts evaluated a selected sample of essays for which human and theoretically driven automated scores differed markedly. Thus, this stage also addressed the question of how the automated approaches compared. Finally, in the fourth stage the automated approaches were applied to the secondary data set containing the main NAEP responses. This stage focused on generalizability, the third research question.

Stage 1

Classroom teachers, state education department staff, and academics expert in the teaching, curricula, assessment, or theory of writing were contacted to participate in the project. Individuals were assigned to one of two committees to create a balance within each committee according to job type and gender.

Each committee consisted of five members (see Appendix C for a list of members). After telephone and email contacts explaining the purposes of the study, each committee met separately for a full day. The day began with a review of the purpose of the study and of approaches to automated essay scoring. Next, both committees reviewed the informative and the persuasive prompts and scoring guides used in the NAEP WOL study and commented on them. Following that, the committee members reviewed e-rater-H's general scoring dimensions and their relations to the NAEP rubrics, again offering critical commentary. Finally, committee members participated in a process for selecting dimension and feature weights.[2]

The weight-selection process used by the two committees differed somewhat in that the first committee was able to make its final selection of features and weights with knowledge of the values empirically derived from the training sample by e-rater-H. Divulging the e-rater-H weights allowed committee members to consider the optimally predictive values (which represent the "operational" behavior of a group of human raters) and their acceptability from a theoretical perspective.[3] This procedure's limitation, of course, is that knowledge of the optimal weights may bias committee judgments away from what they might consider to be more theoretically acceptable values. Because of this fact, the second committee chose its features and weights without knowing anything about the optimally predictive values.

The specific procedures followed by the first committee were:

1. For each essay separately, each member independently assigned initial weights to each of the five e-rater-H writing dimensions (prior to reviewing the specific features that

[2] It is well to note that e-rater v2.1 dimensions are not used in scoring. Rather, features are aggregated directly into a single measure of essay quality. Dimensions were included in this study because the characteristics of good writing are often described at this level of generality and are, therefore, familiar to writing experts and automated essay scoring developers alike. Consequently, an analysis of dimensions offers a good, first-level approximation of the theoretical meaningfulness of an automated program's scores.

[3] Although the rating was done as part of the WOL study, the raters were trained using operational NAEP scoring procedures to make the reading as comparable as possible to production grading.

composed those dimensions). Weights were assigned on a 0-100 scale, with the sum of all dimension weights constrained to equal 100. A weight of zero was to be assigned if the dimension was not relevant to the NAEP scoring guide.

2. The committee was introduced to the specific writing features that e-rater-H computes for each dimension and discussed their meaning and potential relation to the NAEP scoring guide (including identifying features present in the guide but not available in e-rater-H).
3. For essay 1, each member independently assigned initial weights on a 0-100 scale to each feature within each dimension.
4. For essay 1, each member reviewed and readjusted his or her *dimension* weights as needed. This step was taken to allow members to change their view of the relative importance of the dimensions based on their knowledge of the features that e-rater-H uses to mark those dimensions.
5. For essay 1, committee members reviewed the actual feature weights used by e-rater-H and adjusted their own feature and dimension weights as needed.
6. For essay 1, the committee reviewed each member's feature and dimension weights and discussed the differences.
7. For essay 1, each committee member was then given the opportunity to readjust his or her feature and dimension weights based on arguments raised in the discussion.
8. Steps 3 to 7 were repeated for essay 2.
9. For each essay, a single set of final dimension weights was created for the committee by taking the mean weight for each dimension across committee members. Final feature weights were then generated by taking the mean for each feature across committee members and multiplying that mean by the appropriate dimension weight so that the 12 feature weights summed to 100.

The specific procedures followed by the second committee were:

1. For each essay separately, each member independently assigned initial weights to each of the five writing dimensions used by e-rater-H.
2. For each essay separately, the committee reviewed each member's dimension weights and justifications and discussed the differences.
3. For each essay separately, each member readjusted his or her dimension weights based on the arguments raised in the group discussion.
4. The committee was introduced to the specific writing features that e-rater-H computes for each dimension and discussed their meaning and potential relation to the NAEP scoring guide (including identifying features present in the guide but not available in e-rater-H).
5. For essay 1, each member independently assigned initial weights to each feature within each dimension.
6. For essay 1, each member readjusted his or her dimension weights as needed (based on the features composing each dimension).
7. For essay 1, the committee reviewed each member's feature and dimension weights, and discussed the differences.

8. For essay 1, each member readjusted his or her feature and dimension weights based on arguments raised in the discussion.
9. Steps 5-8 were repeated for essay 2.
10. Final dimension and feature weights for committee 2 were created as described for committee 1 above.

Note that in the procedures described above, experts were asked to weight writing dimensions *before* they were introduced to the specific features that composed those dimensions. This procedure separated the perceived theoretical importance of a dimension from its perceived importance *given knowledge of e-rater-H's implementation of it.* Such a separation was desired because e-rater-H's implementation of a dimension may not be what experts mean when they think of that same dimension.

Stage 2

In the second stage, the automated approaches were applied to the WOL data. This stage involved using a training sample of responses to build e-rater models. These models were used to score responses from an independent cross-validation sample (to address research question 2).

The training sample served several purposes. First, it was used by all three approaches to create the vectors of words that are needed for computing feature scores related to topical analysis. Second, it was used for feature weighting and selection. For e-rater-E, this weighting was accomplished through step-wise linear regression, whereas for e-rater-H, hierarchical linear regression was used for 11 of the 12 features. (The weight for the 12th feature, essay length, was set to 30%, a common default used for operational e-rater scoring at that time.) Finally, for all three approaches, the training sample provided the information needed to place e-rater scores on the 1-6 scale used by human raters. This scaling was done somewhat differently for each approach.[4] (See Appendix D for procedural details and Appendix E for an analysis of the impact of these scaling differences on scores.)

The training sample consisted of 250 students selected from the 568 WOL students who had participated in the main NAEP 2002 reading assessment. Because selecting the training sample on the basis of scores on one essay would not necessarily produce a representative distribution of scores on the other essay, 226 of the 250 students were randomly selected proportional to the cross-tabulated score distribution on essays 1 and 2. The remaining 24 students were selected to oversample the tails of the distribution so that there were enough extreme scores to train on. This sampling method was used to ensure that the score distribution covered all score points sufficiently for each essay and was roughly representative of the overall main NAEP reading sample's performance on each essay.

Four e-rater models were built for each essay: one model for e-rater-E, one for e-rater-H, one from the weights derived by the committee that had knowledge of the e-rater-H empirical weights (called "e-rater-T1"), and one based on the weights set by the committee that did not have knowledge of the empirical weights (called "e-rater-T2").

[4] Different automated scoring systems use different scaling procedures because scaling practices were arrived at by different development teams working at different points in time. The scaling differences present in this study are similar in kind to the differences that would result if three commercial automated scoring systems under the control of different companies were used to score the same data set.

The cross-validation sample was composed of those 1,005 essay responses not employed for training. Responses from the cross-validation sample were scored using the parameters derived from the training sample as described in Appendix D.

All responses scored by the original WOL readers were scored by the four e-rater models. Although e-rater includes routines to flag certain types of aberrant responses (e.g., ones that repeat the prompt), these flags were not considered in this study.

Stage 3

In the third stage, experts evaluated a selected sample of essays for which the first human rating and e-rater-T scores differed markedly. This stage also related to research question 2, comparing the automated approaches. For each essay, a sample of 60 responses was selected for which the e-rater-T1 or e-rater-T2 scores diverged most from the human scores awarded to the same responses. The procedure by which these responses were chosen was as follows:

1. All responses were sorted by the size of the score gap between the human score and the e-rater-T score. This procedure was performed separately for e-rater-T1 and e-rater-T2.
2. All responses with a gap > 2 were selected.
3. All responses with gap = 2 were sorted by the human score, producing eight possible groups with human and e-rater-T scores of 1&3, 2&4, 3&5 4&6, 6&4, 5&3, 4&2, and 3&1, respectively.
4. Responses were randomly sampled in turn from each of the groups created in step 3 to meet the target of 60 discrepant responses per prompt, in effect creating a sample of 60 discrepant responses that included *all* study responses with gaps > 2 and a subset of study responses with gaps = 2.

The selected sample of 60 responses per prompt was emailed to the appropriate committee members along with two unlabeled scores, the human score and the e-rater-T score. Committee members were asked to choose the more appropriate score or indicate their own score. In addition, they were asked to justify their choice of score by indicating which factors contributed most to that choice (content, organization, word choice, mechanics, other) and by commenting verbally as appropriate.

Stage 4

In the fourth stage the automated approaches were used to score the main NAEP data to test the generalizability of the theoretically driven model to other essays (research question 3).

The 2002 main NAEP writing assessment included 20 essay prompts. Of those 20 prompts, six were persuasive, seven informative, and seven narrative. Because no narrative prompt was administered in the WOL study, the prompts used to test the generalizability of the theoretically driven approach were drawn from only the persuasive and informative genres (i.e., from among the six persuasive and seven informative prompts used in the main NAEP 2002 assessment). Two essays from each genre were chosen based on two criteria. The first criterion was that the two essays be generally similar to the WOL essay in score distribution. Once this criterion was satisfied, the second consideration related to the task posed by the essay prompt. One essay was selected to be as similar in its task requirements as possible to the WOL prompt and the other essay was selected to be as different as possible from the WOL prompt. Table 6 shows the essay

prompts in terms of three task dimensions: whether stimulus material was provided, the format of the response, and whether the task was abstract or concrete.

TABLE 6

Task Dimensions for WOL Essays and Main NAEP Essays Selected for Evaluating the Generalizability of the Theoretically Driven Approach

Informative			
	WOL Essay 1 "Save a Book"	**Main NAEP Informative Essay 1**	**Main NAEP Informative Essay 2**
Stimulus material	No	No	Yes
Response format	Essay	Essay	Article
Level	Concrete	Abstract	Concrete
Persuasive			
	WOL Essay 2 "School Schedule"	**Main NAEP Persuasive Essay 1**	**Main NAEP Persuasive Essay 2**
Stimulus material	Yes	Yes	No
Response format	Letter	Letter	Essay
Level	Concrete	Concrete	Concrete

The 300 handwritten responses to each of these four prompts were key-entered, with each response verified during key entry. Key entry staff were instructed to preserve spelling, grammatical, and punctuation errors.

Of all the features used by e-rater-E and e-rater-H, only features related to the topical analysis dimension are specific to the prompt. This is because the topical analysis features work by comparing the words used in an essay to the words used in training essays. To generate topical analysis feature scores for responses to the new prompts, a training sample of essays was required to provide the word vectors for each new prompt. This training sample consisted of 100 of the 300 responses for each of the four main NAEP prompts. Responses for the training sample were randomly selected, with oversampling of responses at the top score level (6). The remaining 200 responses were used for the generalizability analysis. This analysis was done using the same automated scoring models (including weights and scaling) for e-rater-E, e-rater-H, e-rater-T1, and e-rater-T2 as originally created for the two essays in the WOL data set.[5]

Analysis

The methods used to analyze the data are described with respect to each of the three research questions.

[5] Note that the special training of Topical analysis features conducted for the generalizability analysis relates only to the computation of raw feature scores. Once computed, these raw feature scores are weighted according to the original scoring model.

To What Extent Are Judgmentally Determined Weights Reproducible?

The reproducibility of judgmentally determined dimension and feature weights was evaluated by assessing the extent of agreement between the two committees. Because the e-rater-H weights are purely statistical and the committee weights are in principle more theoretically based, the committee weights might be expected to be more like one another than like the e-rater-H weights. Consequently, for each dimension, the absolute difference between the two committees' mean weights was compared to the absolute difference between each committee's mean weight and the empirically generated e-rater-H weight. This comparison was done with the initial dimension weights, which were rendered before the committees were exposed to the features used by e-rater-H and before committee 1 reviewed the e-rater-H empirical weights. The comparison was then repeated using the final mean dimension weights and mean feature weights rendered by each committee.

In addition to assessing reproducibility across the two committees, reproducibility was assessed across members within a committee. For this purpose, the range of final weights assigned to each dimension and feature from the e-rater-H set was used.

Finally, the comments of committee members about the relation of the e-rater-H dimensions and features to the NAEP rubrics were summarized. These comments may provide insight into the extent to which e-rater-H's dimensions and features adequately address the target construct. In addition, the comments may give insight into how committee members' weighting choices might vary if e-rater-H had implemented a given dimension differently.

How Do the Approaches to Automated Scoring Compare to One Another in Their Relations to Human Scores and to Other Indicators?

This question was addressed by scoring the same set of essay responses with e-rater-E, e-rater-H, and the two variations of e-rater-T (where each variation of e-rater-T was based on the weights assigned by the appropriate committee). The resulting automated scores were then compared to the human scores for those responses, related to external criteria, and examined to see if they functioned similarly in NAEP reporting groups. It should be noted that comparisons to human scores are unlikely to favor e-rater-T because both e-rater-E and e-rater-H use regression to optimize prediction statistically. However, it is possible that, on cross-validation, e-rater-T will produce prediction that is both almost as good statistically as the empirically based models and more meaningful theoretically (Dawes, 1979). Thus, the main analyses for this question primarily center on determining whether there are practically important differences between e-rater-T and the other approaches to automated scoring. These analyses are summarized in Table 7.

As follow-up to the above analyses, a sample of 60 responses to each of the two essays was analyzed for which the human and e-rater-T1 or e-rater-T2 scores differed markedly. This analysis helps in identifying whether the expert committees find the e-rater-T scores more or less acceptable relative to human graders and, if not, why. To facilitate the analysis, resolved scores for each student response were provided by members of the appropriate expert committee. The analysis included comparing the committee resolved scores, the human scores, and appropriate e-rater-T scores to determine whether the committee resolved scores were more like the human scores or the automated scores. The analysis also included exploring whether the discrepant responses were handled any better by the brute empirical or hybrid approaches than by e-rater-T.

TABLE 7

Analyses Used to Compare the Automated Scoring Approaches

Analysis Question	Cross-Validation Sample	Indices Compared
Does e-rater-T1 or T2 differ from the other automated approaches in its relations to human scores?	255 students (for essay 1) and 242 students (for essay 2) whose responses to each essay have two human scores	Mean scores Machine-human correlations Percentages exact agreement
	1,005 students who responded to both essays	Mean scores Machine-human correlations Percentages exact agreement Inter-prompt correlations
Does e-rater-T1 or T2 differ from the other automated approaches or from human scores in its relations to external indicators?	687 students	Correlation with main NAEP writing plausible values
	318 students	Correlation with main NAEP reading plausible values
	279-1,005 students	Correlations with background and other relevant variables
Does e-rater-T1 or T2 function differently from the other automated approaches in NAEP reporting groups?	1,005 students	Mean scores in each NAEP reporting group

Note. All correlations were Pearson Product-Moment correlations.

How Well Does the Theoretically Driven Scoring Model Developed for One NAEP Prompt Generalize to Other NAEP Prompts of the Same Genre?

To address this question, the e-rater-T scoring model created for grading one essay prompt in each genre was used for scoring two additional prompts from each of the same two genres. Similarly, e-rater-E and e-rater-H models were used to score the responses to each of the four new prompts using the features and weights derived by those programs for evaluating the original WOL prompts.

The generalizability of each scoring approach was evaluated by comparing the different e-rater scores to human scores obtained from the main NAEP data files for those same responses. This comparison was done for each of the four prompts separately. The indices compared included the score means and standard deviations, the Pearson correlations between the human scores and the e-rater scores, and the percentages of exact agreement between the e-rater scores and the human scores. e-rater-T scores should be no different from, and ideally better than, the other approaches in their relations to human scores.

Results

To What Extent Are Judgmentally Determined Weights Reproducible?

The reproducibility of judgmentally determined dimension and feature weights was evaluated across committees and across individuals within a committee. Reproducibility across committees is important if the weights produced by a committee--and the resulting scores--are to have meaning beyond the group of experts that generated them. Reproducibility across members of a committee is also desirable because, to the extent that members agree on the weighting of dimensions and features, their aggregated judgments for each dimension and feature represent a consensus, rather than simply summarizing a more diverse set of views. In the current study, there were only two committees and only five members on each one, so the results with respect to reproducibility should be considered largely as descriptive of what occurred and only suggestive of what might occur from other, similarly conducted weighting activities.

Reproducibility across Committees

How reproducible are the mean weights from one committee to the next? One indication of reproducibility is the absolute difference between the initial mean dimension weights across the two committees, where these initial weights were generated *before* committee members were introduced to either the specific features used by e-rater-H to measure the dimensions or, in the case of committee 1, to the empirical weights employed by e-rater-H. For five dimensions with weights assigned to sum to 100, the mean absolute difference across dimensions will be 40 when the committees totally disagree (i.e., when all of the dimensions given nonzero weight by one committee are given zero weight by the other). On any single dimension, the absolute difference can be as much as 100 points. These worst-case differences, of course, only give a sense of the upper bounds of the scale. For these initial dimension weights, the mean absolute difference between the two committees was 4 points for essay 1 (range of absolute differences for the five dimensions = 1 to 7 points) and 3 points for essay 2 (range = 0 to 7).

Because the e-rater-H weights are purely statistical and the committee weights are in principle more theoretically based, another measure of reproducibility across committees might be the extent to which the committee mean weights are more like one another than they are like the e-rater-H empirically determined ones. To assess reproducibility from this perspective, the mean of the absolute differences between the initial weights assigned to each dimension by committee 1 and committee 2 was compared to the mean of the absolute differences between the weights assigned by each committee and the empirical weights derived by e-rater-H. For both essays, the judgmentally generated means appeared to be considerably closer to one another than to e-rater-H's empirically derived weights. For essay 1, e-rater's mean absolute differences were 20 points with committee 1 and 21 points with committee 2, as compared to the 4-point difference between the two committees. For essay 2, the differences were 17 points with committee 1 and 20 points with committee 2, as compared to 3 points between committees. Table 8 shows the committee and e-rater-H weights.

TABLE 8

Initial Mean Dimension Weights Assigned by Members of Committee 1 and 2, along with e-rater-H Dimension Weights

	Essay 1			Essay 2		
Dimension	**Comm. 1**	**Comm. 2**	**e-rater-H**	**Comm. 1**	**Comm. 2**	**e-rater-H**
Grammar, usage, mechanics, & style	13	16	43	15	15	39
Organization & development	37	36	14	37	38	9
Topical analysis	28	35	6	26	33	12
Word complexity	11	9	8	11	9	10
Essay length	11	4	30	11	5	30

On both essays, e-rater-H gave considerably *higher* weight than either committee to Grammar, usage, mechanics, and style (39% and 43% for e-rater-H vs. 13% to 16% for the committees); and to Essay length (30% for e-rater-H vs. 4% to 11% for the committees). e-rater-H generally gave *lower* weight than either committee to Organization and development and to Topical analysis (20% to 21% for the sum of the two dimensions in e-rater-H vs. 63% to 71% for sum of the two dimensions in either committee). The only dimension on which the empirical and judgmental weights were closely similar was Word complexity.

For all practical purposes, any given dimension in e-rater-H is operationally defined through the specific features used to measure it. Once experts learn how e-rater-H operationally implements its dimensions, those experts may change their dimension weights. Experts may lower their weights, for example, if they view e-rater's implementation of a dimension as insufficient or otherwise inadequate. Alternatively, they may raise their weights if they conclude that an implementation is more complete than expected. Note that raising or lowering the weights for one dimension will necessarily result in a compensatory change in one or more other dimension weights.

Given the changes in expert judgments that may occur as a result of learning about e-rater-H, it is useful to compare the *final* mean dimension weights across committees and also between committees and e-rater-H. These final weights reflect at least a surface-level understanding of the specific features used to indicate each dimension and, for committee 1, how these features were actually weighted by e-rater-H to produce an essay score. (Committee 2 did not see the e-rater-H weights during the weighting process.) Table 9 shows the final mean dimension weights along with the empirical weights used by e-rater-H for each essay.

As the table shows, for the final weights the mean absolute differences between the two committees increased somewhat (from 4 points and 3 points on essays 1 and 2, respectively, to 10 points on each essay). At the same time, the mean absolute difference between e-rater-H and committee 1 (which was shown the empirical weights) became noticeably smaller (from 20 and 17 points for essays 1 and 2, respectively, to 12 and 10 points). In contrast, the mean absolute

difference between e-rater-H and committee 2 (which did not see the empirical weights), decreased by only 1 point for each essay. While far from conclusive, these results suggest that the weight-setting method, in this case sharing vs. not sharing the empirical weights, may affect reproducibility.

TABLE 9

Final Mean Dimension Weights Assigned By Members of Committee 1 and 2 along with e-Rater-H Dimension Weights

	Essay 1			Essay 2		
Dimension	**Comm. 1**	**Comm. 2**	**e-rater-H**	**Comm. 1**	**Comm. 2**	**e-rater-H**
Grammar, usage, mechanics, & style	25	23	43	25	23	39
Organization & development	24	29	14	29	29	9
Topical analysis	19	40	6	16	40	12
Word complexity	13	6	8	11	6	10
Essay length	19	2	30	19	2	30

In terms of specific dimensions, both committees *increased* the weight they assigned to Grammar, usage, mechanics, and style and *decreased* the weights they assigned to Organization and development. These changes perhaps reflect satisfaction or dissatisfaction with the specific features used by e-rater-H to measure these dimensions. Committee weights for two additional dimensions changed, but with the committees moving in opposing directions, perhaps due to the influence on committee 1 of reviewing the empirical weights. Committee 1 decreased its weight for Topical analysis and increased its weight for Essay length, in both cases bringing the judgmental weights closer to the empirically derived ones.

The end result of these changes was that on Topical analysis, committee 1 assigned markedly lower dimension weights than committee 2 (mean weights of 19 vs. 40 and 16 vs. 40). On Essay length, committee 1 assigned notably higher weights than committee 2 (19 vs. 2 for both essays). In comparison to the final committee judgments, e-rater-H gave considerably higher weight than either committee to Grammar, usage, mechanics, and style and to Essay length. e-rater-H generally gave lower weight than either committee to Organization and development and to Topical analysis.

Table 10 gives the mean final *feature* weights for each committee, along with the empirically determined feature weights used by e-rater-H. Table 11 gives the absolute differences between the feature weights. (Essay length appears in Table 10 and 11 as well as in the Tables 8 and 9 because it is both a feature and a dimension.) In combination, these tables give a clearer sense of

TABLE 10

Final Mean Feature Weights Assigned by Members of Committee 1 and 2, along with e-rater-H Feature Weights

Dimension	Feature	Essay 1			Essay 2		
		Comm. 1	**Comm. 2**	**e-rater-H**	**Comm. 1**	**Comm. 2**	**e-rater-H**
Grammar, usage, mechanics, & style	1. Ratio of grammar errors	7	6	17	7	6	19
	2. Ratio of mechanics errors	8	5	11	7	5	9
	3. Ratio of usage errors	6	7	7	6	7	10
	4. Ratio of style errors	3	6	4	5	6	5
Organization & development	5. Number of discourse units	14	6	9	15	7	11
	6. Average length of discourse units	10	23	0	15	23	3
Topical analysis	7. Content similarity with essays in the top score category	9	24	0	5	25	0
	8. The score category containing essays whose words are most similar to the target essay	11	16	12	11	15	6
Word complexity	9. Word repetition	3	2	0	4	2	0
	10. Vocabulary difficulty	7	3	5	4	3	3
	11. Average word length	3	1	5	3	1	5
Essay length	12. Total number of words	19	2	30	19	2	30

Note. All ratios used in feature-score computations are to the total number of words in a response. Each committee feature weight is a mean taken across committee members and rounded to the nearest integer. Except for Essay length, e-rater-H feature weights were produced through a regression procedure and are also rounded to the nearest integer. For e-rater-H, zero represents a regression weight that was either positive but close to zero or negative and subsequently set to zero.

TABLE 11

Absolute Differences for Mean Final Feature Weights between Committees and with e-rater-H

Dimension	Feature	Essay 1			Essay 2		
		Comm. 1 vs. Comm. 2	e-rater-H vs. Comm. 1	e-rater-H vs. Comm. 2	Comm. 1 vs. Comm. 2	e-rater-H vs. Comm. 1	e-rater-H vs. Comm. 2
Grammar, usage, mechanics, & style	1. Ratio of grammar errors	2	10	12	2	11	13
	2. Ratio of mechanics errors	3	4	7	2	2	4
	3. Ratio of usage errors	0	0	0	1	4	3
	4. Ratio of style errors	2	0	2	1	0	1
Organization & development	5. Number of discourse units	9	6	3	8	4	5
	6. Average length of discourse units	14	10	23	8	12	20
Topical analysis	7. Content similarity with essays in the top score category	15	9	24	20	5	25
	8. The score category containing essays whose words are most similar to the target essay	6	2	4	4	6	10
Word complexity	9. Word repetition	1	3	2	2	4	2
	10. Vocabulary difficulty	4	1	2	1	1	0
	11. Average word length	2	2	4	3	2	4
Essay length	12. Total number of words	17	11	28	17	11	28

Note. All ratios used in feature-score computations are to the total number of words in a response. Absolute differences between feature weights were calculated from unrounded values and may not agree exactly with differences calculated from the previous table directly.

how the mean final feature weights differed across committees and from the e-rater-H empirical weights.[6]

As the tables show, when both essays are considered together, the weights generated by the two committees appear to be at least as close to one another as to the e-rater-H weights for four of the 12 features: the Ratio of grammar errors, the Ratio of mechanics errors, the Ratio of usage errors, and Word repetition. The committees consistently diverged more from one another than from the empirical weights for only one feature, the Number of discourse units. The remaining features showed mixed results, often with committee 1 coming closer to the empirical weight than to committee 2. This result occurred consistently for Content similarity with essays in the top score category, the Total number of words (essay length), and the Ratio of style errors (but only marginally for this last feature).

Reproducibility within Committees

How reproducible are weights across individuals within a committee? Tables 12 and 13 give summary statistics for the final *dimension* weights assigned by the individual members of each committee for each of the two essays.

For committee 1 (Table 12), the ranges of the individual member weights were relatively modest except for Essay length, which had a range of weights from 10 to 40 for essay 1 and 10 to 30 for essay 2. For committee 2 (Table 13), the ranges of the weights were substantial for three of the five dimensions: Grammar, usage, mechanics, and style (10-50); Organization and development (0-50); and Topical analysis (20-55).

Table 14 and 15 give summary statistics for the final *feature* weights assigned by each committee for each of the two essays. For committee 1 (Table 14), the only feature for which the range of the individual member weights appeared to be relatively large was Essay length, already noted above in the discussion of dimensions.

For committee 2 (Table 15), three of the 12 features had relatively large ranges. These features were the Average length of discourse units (0-40), Content similarity with essays in the top score category (10-39), and the score category containing essays whose words are most similar to the target essay (10-25).

[6] Note that, because of the manner in which the judgmental weighting process was conducted, committee feature weights are a function of committee dimension weights. That is, giving a low weight to a dimension necessarily results in low weights being assigned to the features that compose that dimension.

TABLE 12

Summary Statistics for Final Dimension Weights for Committee 1

	Essay 1			Essay 2		
	Mean	**SD**	**Range**	**Mean**	**SD**	**Range**
Grammar, usage, mechanics, & style	25	5.0	20-30	25	6.1	20-35
Organization & development	24	4.2	20-30	29	8.9	20-40
Topical analysis	19	5.5	10-25	16	4.2	10-20
Word complexity	13	6.7	5-20	11	5.5	5-20
Essay length	19	12.4	10-40	19	7.4	10-30

TABLE 13

Summary Statistics for Final Dimension Weights for Committee 2

	Essay 1			Essay 2		
	Mean	**SD**	**Range**	**Mean**	**SD**	**Range**
Grammar, usage, mechanics, & style	23	15.7	10-50	23	15.7	10-50
Organization & development	29	19.5	0-50	29	19.5	0-50
Topical analysis	40	13.7	20-55	40	13.7	20-55
Word complexity	6	4.0	0-10	6	4.0	0-10
Essay length	2	2.6	0-5	2	2.6	0-5

TABLE 14

Summary Statistics for Final Feature Weights for Committee 1

		Essay 1			Essay 2		
Dimension	**Feature**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Grammar, usage, mechanics, & style	1. Ratio of grammar errors	7	3.6	2-10	7	4.5	2-14
	2. Ratio of mechanics errors	8	1.9	5-10	7	2.3	5-11
	3. Ratio of usage errors	6	3.2	2-10	6	3.1	2-11
	4. Ratio of style errors	3	3.2	0-8	5	3.4	0-8
Organization & development	5. Number of discourse units	14	3.1	10-18	15	4.5	10-20
	6. Average length of discourse units	10	3.4	4-13	15	4.5	10-20
Topical analysis	7. Content similarity with essays in the top score category	9	4.9	0-13	5	4.7	0-10
	8. The score category containing essays whose words are most similar to the target essay	11	1.1	10-13	11	5.2	8-20
Word complexity	9. Word repetition	3	2.8	1-7	4	2.3	1-7
	10. Vocabulary difficulty	7	5.5	2-16	4	1.8	2-7
	11. Average word length	3	2.1	2-7	3	2.1	2-7
Essay length	12. Total number of words	19	12.4	10-40	19	7.4	10-30

Note. All ratios are to the total number of words in a response.

TABLE 15

Summary Statistics for Final Feature Weights for Committee 2

		Essay 1			Essay 2		
Dimension	**Feature**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Grammar, usage, mechanics, & style	1. Ratio of grammar errors	6	4.3	2-13	6	4.3	2-13
	2. Ratio of mechanics errors	5	4.4	2-13	5	4.4	2-13
	3. Ratio of usage errors	7	3.9	3-13	7	3.9	3-13
	4. Ratio of style errors	6	4.5	3-13	6	4.5	3-13
Organization & development	5. Number of discourse units	6	4.6	0-10	7	5.1	0-11
	6. Average length of discourse units	23	15.1	0-40	23	14.9	0-40
Topical analysis	7. Content similarity with essays in the top score category	24	10.3	10-39	25	10.6	10-39
	8. The score category containing essays whose words are most similar to the target essay	16	6.4	10-25	15	6.2	10-25
Word complexity	9. Word repetition	2	1.6	0-4	2	1.4	0-4
	10. Vocabulary difficulty	3	1.9	0-5	3	2.0	0-6
	11. Average word length	1	0.7	0-2	1	0.7	0-2
Essay length	12. Total number of words	2	2.6	0-5	2	2.6	0-5

Note. All ratios are to the total number of words in a response.

Qualitative Judgments

As part of the weighting process, committee members were asked to judge qualitatively the extent to which the e-rater-H dimensions and features adequately represented the NAEP rubrics. As a preliminary observation, both committees noted that the NAEP persuasive and informative rubrics differed from one another only in a single requirement. For the persuasive essay, that requirement was to take a clear position and develop it. This minimal difference was cited by members as the reason for the close similarity in committee weights across the two essays.

With respect to the first dimension, Grammar, usage, mechanics, and style, it was noted that style was missing from both the scoring rubrics and, in large part, from e-rater-H. The e-rater-H style feature counts as errors such things as repetition, overuse of short or of long sentences, and the use of passive voice. This implementation was viewed as missing the essence of style as embodied in such text characteristics as extended metaphor, personal voice, figurative language, rhetorical devices (e.g., purposeful repetition), language sophistication, and unconventional organization. Committee members also expressed the concern that some aspects of style might be mistakenly treated as errors (e.g., the short sentences that characterized Hemingway's style or the repetitions that marked King's "I Have a Dream" speech). Finally, for persuasive essays, e-rater-H does not detect the use of different types of rhetorical style (ethos, pathos, and logos).

Regarding Organization and development, committee members viewed e-rater-H's representation of this dimension as too limited because the five-paragraph model (introduction, three main ideas, summary) was the only acceptable organizational scheme. Experts viewed this model as encouraging a "superficial, template-based approach to writing." They noted that, for an informative essay, a single, well-developed main idea might suffice, while for a persuasive essay, the requirement for three main ideas was more justified. Rather than conceptualizing organization in terms of the five-paragraph model, one committee member suggested shifting to a conception based on claims and evidence. Instead of counting the number of discourse units, a more theoretically meaningful approach to evaluating an essay's depth and quality of elaboration would be to look for the hierarchal structure of evidence supporting each claim.

Committee members also thought that audience awareness was missing from both e-rater-H's implementation of Organization and development and from the NAEP rubrics. Audience awareness, they suggested, might be detected through particular key words or phrases commonly found in high-scoring essays. Finally, the committees noted that cohesion was implied by the NAEP rubrics' inclusion of transitions but that e-rater-H appeared to take no explicit account of cohesion through its Organization and development features (e.g., the presence of logical connectors like "if," "then," and "because").

The third e-rater-H dimension was Word complexity. Committee members noted that "word choice" was included in the NAEP rubrics. The experts' view was that "choice" was a more appropriate consideration than "complexity" because more difficult words are not necessarily better ones. Both committees gave this dimension a relatively low weight.

Regarding Topical analysis, the fourth e-rater-H dimension, members observed that this characteristic was more explicit in e-rater-H than in the NAEP rubrics, which in their view gave insufficient attention to content or to the quality of ideas, especially for the informative essay.

Finally, the experts noted that essay length was measured by e-rater but was not included in the NAEP rubrics explicitly.

As should be evident from the above description, committee members felt important dimensions were either missing from, or too narrowly represented by, e-rater-H's features (and sometimes also from the NAEP rubrics). As a consequence, those members might well have assigned different dimension weights had the representation of these dimensions and features been more in agreement with their views on good writing. Indeed, some members said that the weighting assignment was very difficult because giving low or no weight to dimensions they felt were inadequately measured by e-rater-H inevitably resulted in overweighting other dimensions.

Different committee members might have resolved this classic "avoidance-avoidance" conflict in different ways.

How Do the Approaches to Automated Scoring Compare to One Another in Their Relations to Human Scores and to Other Indicators?

This question was addressed by scoring the same set of essay responses with e-rater-E, e-rater-H, and two variations of e-rater-T. Four categories of analysis were run. These analyses concerned relations with human scores, relations with other indicators, functioning in NAEP reporting groups, and resolution of large machine-human score discrepancies.

Relations with Human Scores

As part of the NAEP WOL study, two groups of human raters scored typed responses presented to them onscreen, with each essay scored by a different group of raters. A random sample of approximately 25% of the responses was scored by a second rater in each group. Table 16 gives the mean scores for human ratings and for the automated scoring approaches. Results are given for the full cross-validation sample of 1,005 students and for the subsample having two human scores.

Several analyses were done using the scores summarized in the table. First, for the subsample with two human scores, the difference between these two scores was tested. That test showed no significant difference between the first and second human scores for essay 1 ($t_{254} = -1.07$, $p>.05$) or for essay 2 ($t_{241} = 1.51$, $p>.05$), suggesting that the two human ratings could be considered to have come from the same population. As a consequence, the two human scores were averaged to form a more reliable estimate of each examinee's true score. Only that average score (labeled "Human R1 + R2") is given in the table for the subsample.

Next, in the subsample with two human scores, a repeated-measures ANOVA was executed to test the difference between the mean scores produced by the five methods (one combined human rater and four automated raters). This ANOVA was applied separately for each essay, with scoring method as the independent variable and essay score as the dependent variable. A significant effect was found for scoring method for essay 1 ($F_{4,1016} = 8.2$, $p < .001$) and for essay 2 ($F_{4,964} = 10.0$, $p < .001$). Post-hoc tests contrasting each automated score against the combined human score indicated that the e-rater-T2 score was significantly lower than the combined human score for both essay 1 ($F_{1,254} = 14.9$, $p < .001$) and essay 2 ($F_{1,241} = 18.5$, $p < .001$). The effect sizes were small: .20 and .21 standard deviation units for essay 1 and essay 2, respectively. In addition, e-rater-T1 produced significantly lower scores than the combined human score for essay 2 ($F_{1,241}) = 5.7$, $p < .05$), with an effect size of .11.

The above analysis was repeated in the full cross-validation sample (N = 1,005), with the human method represented only by the first rating. Once again significant effects were found for scoring method on both essays (for essay 1, $F_{4,4016} = 28.5$, $p < .001$ and for essay 2, $F_{4,4016} = 40.1$, $p < .001$). However, post-hoc tests showed that more of the machine methods differed from the human method. For essay 1, e-rater-H awarded scores that were significantly *higher* than the scores given by the first human rating (effect size = -.06), while e-rater-T2 awarded scores that were significantly lower than that first human rating (effect size =.16). For essay 2, *all* machine methods produced scores that were significantly *lower* than the human scores (effect size range = .06 to .24). (See Appendix F for post-hoc test results and effect sizes.)

TABLE 16

Summary Statistics for Essay Scores in the Total Cross-Validation Sample (N = 1,005) and in the Cross-Validation Subsample Scored by Two Human Raters (N = 255/242)

Scoring Method	**Mean**	**SD**	**Mean**	**SD**
Essay 1	**N = 1,005**		**N = 255**	
Human R1	3.6	1.2	-	-
Human R1 + R2	-	-	3.7	1.1
e-rater-E	3.6	1.0	3.7	1.0
e-rater-H	3.7	1.3	3.7	1.2
e-rater-T1	3.6	1.3	3.6	1.2
e-rater-T2	3.4	1.2	3.4	1.3
Essay 2	**N = 1,005**		**N = 242**	
Human R1	3.5	1.2	-	-
Human R1 + R2	-	-	3.5	1.2
e-rater-E	3.4	1.0	3.4	0.9
e-rater-H	3.4	1.3	3.5	1.3
e-rater-T1	3.3	1.2	3.3	1.2
e-rater-T2	3.2	1.3	3.2	1.2

Note. Human R1 = first human rating. Human R1 + R2 = the mean of the two human ratings.

Table 17 shows the intercorrelations among the four automated scoring approaches. Of note is that the e-rater-T1 and e-rater-T2 approaches strongly intercorrelated (r = .86 for essay 1 and .90 for essay 2). Even so, the two methods were significantly different in their relations to the other automated approaches. e-rater-T1's correlation with e-rater-H was significantly higher than e-rater-T2's correlation with e-rater-H for essay 1 (.92 vs. 81.) (t_{1002} = 17.14, p < .01) as well as for essay 2 (.90 vs. .84) (t_{1002} = 10.87, p < .01). Also, e-rater-T1's correlation with e-rater-E was significantly higher than e-rater-T2's correlation with e-rater-E for essay 1 (.77 vs. .67) (t_{1002} = 9.35, p < .01) as well as for essay 2 (.81 vs. .74) (t_{1002} = 8.36, p < .01). These differences in functioning between the two e-rater-T approaches can only be due to the feature weights, which constitute the sole distinction between them.

TABLE 17

Intercorrelations among the Automated Essay Scoring Approaches for the Total Cross-Validation Sample (N = 1,055)

	e-rater-E	**e-rater-H**	**e-rater-T1**
Essay 1			
e-rater-H	.75	-	
e-rater-T1	.77	.92	-
e-rater-T2	.67	.81	.86
Essay 2			
e-rater-H	.77	-	
e-rater-T1	.81	.90	-
e-rater-T2	.74	.84	.90

Note. All correlations are significantly different from zero at $p < .05$.

Table 18 shows the percentages of exact agreement among the four automated approaches. Kappa (Fleiss, 1981) is also given. The pattern of results is similar to that depicted by the correlations. e-rater-T1 and T2 agreed in a majority of cases with one another (59% of the time for essay 1 and 69% for essay 2) but behaved differently from one another vis-à-vis e-rater-H. For essay 1, T1 agreed with e-rater-H 74% of the time, whereas e-rater-T2 agreed with e-rater-H in only 48% of cases. For essay 2, T1 agreed with e-rater-H in 70% of instances. The comparable percentage for e-rater-T2 was 52%.

Table 19 gives the correlations of each e-rater approach with the first human rating and with the mean of the two human ratings. Because the e-rater-E and e-rater-H feature weights were selected to optimally predict the scores awarded by these same human raters in the training sample, the e-rater-T1 and T2 scores should not be expected to agree with the human scores more highly than the empirically based methods. Interestingly, the correlations between e-rater-T1 and the human scores were virtually identical to those between e-rater-H and the human scores for both essays. Further, for essay 2, the T1 scores correlated significantly higher with the human scores than the e-rater-E scores correlated with the human scores ($t_{1002} = 3.46$, $p < .01$ for the first human score and $t_{1003} = 2.60$, $p < .01$ for the mean of the human scores). In contrast, the e-rater-T2 scores correlated consistently less well with humans than did the e-rater-H scores (for essay 1, $t_{1002} = -6.03$, $p < .01$ for the first human score and $t_{1003} = -5.02$, $p < .01$ for the mean of the human scores; for essay 2, $t_{1002} = -3.07$, $p < .01$ for the first human score and $t_{1003} = -4.57$, $p < .01$ for the mean of the human scores). The e-rater-T2 scores also had correlations with human scores that were lower than the correlations of e-rater-E with human scores, but only for essay 1 ($t_{1002} = -3.94$, $p < .01$ for the first human score and $t_{1003} = -3.30$, $p < .01$ for the mean of the human scores). The differences in functioning between the two versions of e-rater-T derive from their feature weights. For T1, these weights were closer to the optimal, empirically derived weights used by e-rater-H.

TABLE 18

Percentage Exact Agreement (and Kappa) among the Automated Scoring Models for the Total Cross-Validation Sample (N = 1,005)

	e-rater-E	**e-rater-H**	**e-rater-T1**
Essay 1			
e-rater-H	50 (.34)	-	
e-rater-T1	51 (.35)	74 (.67)	-
e-rater-T2	41 (.22)	48 (.33)	59 (.47)
Essay 2			
e-rater-H	52 (.34)		
e-rater-T1	53 (.36)	70 (.60)	
e-rater-T2	45 (.27)	52 (.38)	69 (.60)

Note. Kappa appears in parentheses. All kappa values are significantly different from zero at $p < .05$.

TABLE 19

Correlations of the Automated Essay Scoring Approaches with Human Ratings for the Total Cross-Validation Sample (N = 1,055) and for Students in the Cross-Validation Sample Whose Essays Were Scored by Two Human Raters (N = 255/242)

	e-rater-E	**e-rater-H**	**e-rater-T1**	**e-rater-T2**
Essay 1				
Human R1	.66	.67	.66	.59
Human R1 + R2[a]	.72	.73	.74	.67
Essay 2				
Human R1	.69	.72	.73	.68
Human R1 + R2[a]	.72	.75	.75	.70

[a] Correlations with Human R1 + R2 are based on N = 255 participants for essay 1 and on 242 participants for essay 2. The correlation between the two human ratings was .78 for essay 1 and .87 for essay 2.

Note. Human R1 = first human rating. Human R1 + R2 = the mean of the two human ratings.

In Table 20, the percentages of exact agreement between each automated approach and the human ratings are given. For this analysis, agreement with the first and with the second human ratings is presented separately because averaging the two human ratings often does not produce an integer score. Here, e-rater-T1's exact agreement was between 3 and 7 points lower in these samples than was e-rater-H's agreement and 2 to 6 points lower than e-rater-E's agreement with the human ratings. e-rater-T2's agreement ran between 6 and 14 points lower than e-rater-H's values and between 8 and 12 points lower than the e-rater-E's values. e-rater-T1's exact agreement was higher than e-rater-T2's exact agreement by 4 to 8 percentage points.

TABLE 20

Percentage Agreement (and Kappa) of the Automated Essay Scoring Approaches with Human Ratings for the Total Cross-Validation Sample (N = 1,055) and for Students in the Cross-Validation Sample Whose Essays Were Scored by Two Human Raters (N = 255/242)

	e-rater-E	**e-rater-H**	**e-rater-T1**	**e-rater-T2**
Essay 1				
Human R1	43 (.24)	41 (.24)	39 (.21)	35 (.16)
Human R2	49 (.30)	49 (.33)	43 (.26)	37 (.18)
Essay 2				
Human R1	48 (.30)	50 (.35)	44 (.28)	37 (.18)
Human R2	48 ([*])	51 (.36)	46 (.30)	39 (.22)

[*]Kappa could not be calculated due to absence of values for score level "1" for e-rater-E.

[a] Correlations with Human R1 + R2 are based on N = 255 participants for essay 1 and on 242 participants for essay 2. The percentage of agreement (kappa) between the two human ratings was 59 (.46) for essay 1 and 63 (.52) for essay 2.

Note. Kappa appears in parentheses. Human R1 = first human rating. Human R2 = second human rating.

The last analysis in this section compares, for each of the four automated scoring approaches, the correlation between the two essay prompts with the same correlation computed from human scores. This analysis uses the total cross-validation sample. Table 21 gives the results. As the table shows, the correlation between scores on the two essays as assigned by the first human rating was .61. The methods with correlations significantly different from this value were e-rater-E ($t_{1002} = 3.24$, $p < .01$) and e-rater-T2 ($t_{1002} = 2.77$, $p < .01$), both of which had cross-essay correlations lower than the human value.

TABLE 21

Correlations between Essay 1 and Essay 2 Scores for the Total Cross-Validation Sample (N = 1,055)

Scoring Method	Correlation between Essays
Human R1	.61
e-rater-E	.54*
e-rater-H	.64
e-rater-T1	.63
e-rater-T2	.55*

* Correlation is significantly different from correlation for Human R1 at $p < .05$.

Note. Human R1 = first human rating.

Relations with Other Indicators

Whereas the above analyses centered on comparison of the automated methods with human scores, the analyses in this section explore the extent to which the different automated methods can be distinguished in their relationships to other indicators. Among the measures in the WOL data set were indicators of typing speed, self-reported activities related to writing and to reading, main NAEP writing plausible values, main NAEP reading plausible values, and the number of words comprising each essay.

Each of these variables deserves comment. The first variable, typing speed, was the number of words entered within two minutes from a 78-word passage, not accounting for accuracy errors. Typing speed is of interest because it is conceivable that students with different levels of typing proficiency will produce essays of different lengths. To the extent that the automated methods are more or less sensitive to length as compared with human raters, faster typists could be graded differently depending upon the scoring method. The second and third variables, related to writing and reading activities, were computed from student responses to questionnaires. Within each set of questions (i.e., reading or writing), the number of response categories varied. Each question was therefore rescaled by taking the mean of the item responses to that question, subtracting the minimum possible value, and dividing by the number of scale points. These scores were then summed across all items separately in the reading activities and writing activities sets. The standardized coefficient alpha for the 30 questions comprising the writing activities scale was .82. For the 26 questions in the reading activities scale, it was .86.

Main NAEP writing plausible values or main NAEP reading plausible values were available for different subsets of students in the cross-validation sample, depending upon which of the two main NAEP assessments a student had previously taken. These plausible values represent five random draws from an estimated ability distribution based upon student responses to the test (writing or reading), demographic information, and estimated item parameters. All five draws are used (independently) in conducting any given analysis. Of particular importance to the current study is that the writing plausible values generated from main NAEP were computed

from a *different* pair of essay prompts than the ones scored by the automated methods. Also, the human graders used to score those prompts were different from the ones employed in the analyses presented above. Finally, the scores awarded by the human graders are moderated through the plausible values methodology, producing a more accurate estimate of group means than would otherwise be obtained (see Allen, Donoghue, & Schoeps, 2001, for a description of the methodology used to generate plausible values.)

Among the indicators included in this section, the ones with the clearest relevance are the writing activities and writing plausible values. Finding that a given automated approach is more highly related to these variables than are other automated approaches would argue for the validity of that approach. The interpretation of relationships with reading activities and reading plausible values is less straightforward. That is, it might not necessarily be the case that the more valid automated essay scoring approach is the one with the highest relationship to reading performance or reading activities. However, any difference in relations with these indicators would still suggest that the automated methods are not functioning equivalently, so such relationships are included here. Finally, even though essay length is explicitly represented in e-rater-H's and e-rater-T's scoring, how this characteristic relates to the scores ultimately produced by these approaches is unclear. This uncertainty stems from the fact that length is implicitly represented through other features (e.g., the product of the Number of discourse units and the Average length of discourse units is equivalent to essay length). Any differences observed among the methods in their relations with essay length would also suggest distinctions in the meaning of their scores.

Shown in Table 22 are the correlations between each scoring approach and the other indicators (see Appendix G for summary statistics for these indicators). Several findings were consistent across the two essays when the external relations of the automated approaches were contrasted with those of the first human rating. First, e-rater-E's correlation with main NAEP writing performance (represented by the plausible values) was significantly lower than the correlation between the first human rating and main NAEP performance ($t_{684} = 2.40$, $p < .05$ for essay 1 and $t_{684} = 3.78$, $p < .01$ for essay 2). Second, e-rater-T2's correlation was lower with main NAEP reading performance (represented by the plausible values) than was the first human rating's correlation with that performance ($t_{315} = 2.18$, $p < .05$ for essay 1 and $t_{315} = 2.12$, $p < .05$ for essay 2). Finally, *all* of the automated methods correlated more strongly with essay length than did the first human rating (t_{1002} range = -8.60 to -23.06, $p < .01$ for essay 1 and t_{1002} range = -6. 84 to -19.22, $p < .01$ for essay 2).

With respect to essay length, e-rater-T1 was more related to this feature than was e-rater-H (t_{1002} = -9.84, p < .01 for essay 1 and t_{1002} = -5.17, p < .01 for essay 2). This higher relationship occurred even though e-rater-T1's length feature weight was 19% as compared with 30% for e-rater-H. (This result appears to have occurred because of the higher weight given by e-rater-T1 to the two Organization and development features which, together, largely duplicate Essay length.) e-rater-H was, in turn, more related to length than was either e-rater-E ($t_{1002} = 5.06$, p < .01 for essay 1 and $t_{1002} = 3.18$, p < .01 for essay 2) or e-rater-T2 ($t_{1002} = 6.66$, p < .01 for essay 1 and $t_{1002} = 7.95$, p < .01 for essay 2).

TABLE 22

Correlations between the Scoring Approaches and Other Indicators for the Cross-Validation Sample

Essay 1						
Indicator	**N**	**Human R1**	**e-rater-E**	**e-rater-H**	**e-rater-T1**	**e-rater-T2**
Typing speed	948	.44	.43	.47	.48	.44
Writing activities	629	.11	.12	.16	.13	.13
Reading activities	279	-.05	.09*	.08*	.14*	.14*
Main NAEP writing[a]	687	.52	.46*	.49	.49	.44*
Main NAEP reading[a]	318	.48	.42	.47	.45	.38*
Essay length	1,005	.57	.74*	.81*	.87*	.73*
Essay 2						
Typing speed	948	.49	.47	.47	.52	.51
Writing activities	629	.12	.14	.21*	.19*	.19*
Reading activities	279	-.04	.05	.04	.02	.00
Main NAEP writing[a]	687	.56	.47*	.53	.53	.52
Main NAEP reading[a]	318	.53	.46	.45*	.45*	.46*
Essay length	1,005	.66	.81*	.84*	.87*	.76*

* Correlation significantly different from the correlation of the first human rating with the relevant indicator at ($p < .05$).

a The correlations reported are averaged across five plausible values using the Z-score transformation.

Note. Human R1 = first human rating.

Functioning in NAEP Reporting Groups

This analysis focuses on the functioning of the automated methods within NAEP reporting groups, in particular groups defined by race/ethnicity, gender, parents' education level, eligibility for free or reduced price school lunch (an indicator of socioeconomic status), and school location. Table 23 gives the distribution of these characteristics for the total cross-validation sample.

To analyze the extent to which the different automated approaches functioned similarly across the NAEP reporting groups, a repeated-measures ANOVA was conducted separately for each essay and for each reporting group. The independent variables were the reporting group of interest (e.g., gender) and scoring method (four automated approaches and the first human score), with repeated measures on scoring method. The dependent variable was the essay score. Of particular interest in this analysis was whether there is a significant group-by-scoring-method

TABLE 23

Distribution of Students within NAEP Reporting Groups for the Total Cross-Validation Sample (N = 1,005)

NAEP Reporting Group	Category	N	Percent
Race/ethnicity	White	672	67
	Black	168	17
	Hispanic	110	11
	Asian	32	3
	American Indian	13	1
	Unspecified	10	1
Sex	Male	529	53
	Female	473	47
	Unspecified	3	0
Parents' education level	Less than HS	53	5
	Graduated HS	178	18
	Some education after HS	215	21
	Graduated college	466	46
	Unspecified	93	9
Eligibility for free or reduced-price school lunch	Eligible[a]	287	29
	Not eligible	600	60
	Unspecified	118	12
School location	Central city	252	25
	Urban fringe/large town	394	39
	Rural	359	36

[a] This group includes students eligible for free lunch and those eligible for reduced-price lunch.

interaction, suggesting the possibility that the automated approaches differ from one another in the mean scores they assign to particular groups.

For this analysis, some groups were dropped and others combined. For all analyses, the unspecified group was dropped. For race/ethnicity, the Asian and American Indian groups were not included because of their small sample sizes. For parents' education level, the less-than-high-school and graduated-high-school groups were combined to create a group with high school degree or less. The some-education-after-high-school and graduated-college groups were also collapsed to form a group, more-than-high-school-degree.

Table 24 gives the ANOVA results. As the table indicates, the group-by-scoring-method interaction was significant for both essays in three of five instances: race/ethnicity, sex, and eligibility for free or reduced-price school lunch. This result suggests that the differences between categories within each of these reporting groups are not the same across scoring methods.

To identify which method(s) operated differently in each of these three reporting groups, a repeated-measures ANOVA was run separately for each level of the group variable (e.g., a separate ANOVA for males and one for females) for each essay. Scoring method was the independent variable and essay score was the dependent variable. Results showed scoring method to be significant for each level of the group variable in each analysis, indicating that one or more of the scoring methods operated differently from the other scoring methods for each group. Post-hoc contrasts comparing each automated method with the first human score were executed next to try to identify the groups and scoring methods.

The post-hoc contrasts suggested that, relative to the first human rating, the interaction between scoring method and NAEP reporting group was usually not consistent across essays for any given automated method (see Appendix H for the contrasts and effect sizes). Further, the effect sizes were generally small and often inconsequential, with the largest effect being .29 standard deviation units between the first human rating and e-rater-T2 for White students on essay 2. The only effects that appear to show evidence of a consistent scoring-method-by-reporting-group interaction were for gender. Here, e-rater-T1 and e-rater-T2 consistently awarded scores to males that were lower than the first human rating.

Resolution of Large Human-Machine Score Discrepancies

As follow-up to the above analyses, a sample of 60 responses to each of the two essays was analyzed for which the human and e-rater-T scores differed markedly. To help identify whether the expert committees found the e-rater-T scores more or less acceptable relative to human scores, each committee member was sent by email the discrepant responses resulting from the application of e-rater-T with that committee's weights. Committee members were also given the first human rating and the e-rater-T scores. In this sample of discrepant responses, the percentage of instances in which the human score was higher than the e-rater-T score was, for e-rater-T1, 40% for essay 1 and 45% for essay 2. For e-rater-T2, the percentages were 50% and 48% for essays 1 and 2, respectively. For each discrepant response, committee members were asked to choose blindly the more appropriate score (human or e-rater-T) or indicate their own score. Members made their judgments individually and not as a committee. Four members from committee 1 and five from committee 2 returned resolved scores.

TABLE 24

ANOVA Results Comparing the Functioning of Scoring Approaches in NAEP Reporting Groups for the Total Cross-Validation Sample

		Essay 1		Essay 2	
NAEP Reporting Group	**Effect**	**F**	**P**	**F**	**P**
Race/ethnicity	Race/ethnicity	27.8 (2,947)	.001	28.7 (2,947)	.001
	Scoring method	14.0 (4,3788)	.001	22.2 (4,3788)	.001
	Race/ethnicity x scoring method	3.6 (8,3788)	.001	5.7 (8,3788)	.001
Sex	Sex	54.5 (1,1000)	.002	50.6 (1,1000)	.001
	Scoring method	4.2 (4,4000)	.001	50.2 (4,4000)	.001
	Sex × scoring method	7.4 (4,4000)	.001	6.5 (4,4000)	.001
Parents' education level	Parents education level	48.5 (1,910)	.003	128.1 (1,910)	.001
	Scoring method	13.1 (4,3640)	.001	25.8 (4,3640)	.001
	Parents' education level × scoring method	2.2 (4,3640)	.060	1.0 (4,3640)	.41
Eligibility for free or reduced-price lunch	Eligibility	47.1 (1,885)	.001	44.1 (1,885)	.001
	Scoring method	4.3 (4,3540)	.002	34.4 (4,3540)	.001
	Eligibility × scoring method	8.9 (4,3540)	.001	3.6 (4,3540)	.006
School location	School location	4.8 (2,1002)	.009	2.3 (2,1002)	.090
	Scoring method	4.1 (4,4008)	.003	49.6 (4,4008)	.001
	School location × scoring method	0.2 (8,4008)	1.0	0.4 (8,4008)	.910

Note. Race had three levels (White, Black, Hispanic), sex had two levels, parents' education level had two levels (high school degree or less, more than high school degree), eligibility for free or reduced-price school lunch had two levels (eligible, not eligible), and school location had three levels (central city, urban fringe/large town, rural).

Table 25 shows summary statistics for the resolved scores, the human scores, and the scores awarded by each of the automated approaches. Results of a statistical test of the differences

among the five mean scores (a resolved score, the first human, and three automated scores) are also indicated. The statistical test was a repeated-measures ANOVA conducted separately for each essay and version of e-rater-T, with scoring method as the independent variable and essay score as the dependent variable. The data are relevant to how accurate the e-rater-T scores are for this sample of discrepant responses, as well as whether the other automated scoring approaches produce more accurate scores than e-rater-T. (With respect to this second issue, however, these data need to be viewed cautiously as the included responses were chosen because e-rater-T--and not the other approaches--scored them discrepantly.)

As the table indicates, the effect for scoring method was significant in all four samples. Post-hoc contrasts were conducted against the first human rating because that rating best represented the NAEP scale on which the automated approaches were intended to report. These contrasts showed that the mean resolved score was *always* significantly lower than the first human score, suggesting that the experts consistently held to a higher standard than the NAEP raters. Further, in only one sample (i.e., for committee 1 on essay 1), was the e-rater-T mean significantly different from the first human mean. In that instance, *all* of the automated approaches produced scores that were significantly higher than the first human score (which, as noted, was itself significantly higher than the resolved score). For two other samples, the automated scores were not significantly different from the first human score. For the last sample (committee 2 on essay 2), e-rater-H produced significantly higher scores than the first human score.

TABLE 25

Summary Statistics and ANOVA Results for the Resolved Scores of Committee Members, the First Human Rating, and the Scores from the Automated Approaches

		Human R1	**e-rater-E**	**e-rater-H**	**e-rater-T1**	**Mean Resolved Score**	$F_{(4,236)}$	**P**
Committee 1								
Essay 1	Mean	3.2	3.7*	3.8*	3.8*	2.8*	13.3	.001
	SD	1.5	1.1	1.5	1.6	1.2		
Essay 2	Mean	3.2	3.5	3.7	3.6	2.6*	12.1	.001
	SD	1.6	1.1	1.4	1.5	1.3		
Committee 2								
Essay 1	Mean	3.5	3.6	3.6	3.5	3.0*	3.2	.01
	SD	1.6	.9	1.3	1.6	1.1		
Essay 2	Mean	3.3	3.5	3.8*	3.6	3.1*	4.7	.01
	SD	1.7	1.1	1.3	1.5	1.3		

*Significantly different from Human R1 score at $p < .05$.

Note. A separate sample of 60 responses was selected for each committee and essay. Committee 1 reviewed discrepant responses for e-rater-T1 and committee 2 reviewed discrepant responses for e-rater-T2. Human R1 = first human rating.

Although there appeared to be little distinction among the automated approaches in the mean scores awarded to discrepant responses, the rank order of their scores could well be different. Table 26 gives the correlations between the mean resolved scores and each of the scoring methods. In three of the four samples, the mean resolved scores correlated significantly higher with the first human score than with any of the automated scores, suggesting that the human scores are more credible indictors of proficiency than the automated methods (t_{57} range = 2.86 to 11.64, $p < .05$). For these three samples, the differences between the human and machine correlations were, in practical terms, very substantial, with the *smallest* difference in each sample running between 18 and 23 points.

TABLE 26

Correlations between Mean Resolved Scores of Committee Members and Automated Essay Scoring Approaches

	Human R1	**e-rater-E**	**e-rater-H**	**e-rater-T**
Committee 1				
Essay 1	.71	.72	.67	.58
Essay 2	.80	.60*	.46*	.52*
Committee 2				
Essay 1	.80	.58*	.62*	.28*
Essay 2	.86	.63*	.55*	.36*

*Significantly different from the correlation of Human R1 and resolved score at $p < .05$.

Note. Human R1 = first human rating. A separate sample of 60 responses was selected for each committee and essay. Committee 1 reviewed discrepant responses for e-rater-T1 and committee 2 reviewed discrepant responses for e-rater-T2.

There were also differences among the automated approaches in their relations with the resolved scores. For both essays, the e-rater-H scores correlated higher with the resolved scores than the e-rater-T2 scores correlated with the resolved scores ($t_{57} = 6.34$, $p < .05$ for essay 1 and to $t_{57} = 3.10$, $p < .05$ for essay 2). And, for both essays, the e-rater-E scores correlated higher with the resolved scores than did the e-rater-T2 scores ($t_{57} = 3.55$, $p < .05$ for essay 1 and to $t_{57} = 3.67$, $p < .05$ for essay 2). Last, for essay 1 the e-rater-H scores correlated significantly higher with the resolved scores than did the e-rater-T1 scores ($t_{57} = 2.34$, $p < .05$).

To get a better understanding of the factors that might have influenced committee members in choosing their resolved scores, members were asked to check one or more of five categories: Content, Organization, Word choice, Mechanics, Other (e.g., style, audience). The number of instances in which each category was selected was summed across all members of a committee and all responses to a prompt to suggest the importance of the category in determining the resolved score. These sums were tabulated separately for the cases in which the mean resolved score agreed more closely with the first human score, agreed more closely with the e-rater-T

score, or was exactly in between. The results are suggestive only, as reasons were not given by all committee members.

Table 27 shows the results in terms of the percentages of the total number of reasons given. For all three "gap type" categories, the primary reasons indicated by committee members for choosing a resolved score were based on the content of the essay and its organization. The remaining three categories were of secondary or, sometimes, negligible importance. For the subsample of responses resolved in favor of the first human rating, 70% of the offered reasons fell into the Content or Organization categories. For the subsample resolved in favor of e-rater-T, the comparable figure was 76%, while for the subsample where the resolved scores fell in between, it was 69%. Also, in these subsamples, the percentage of reasons falling into Content was usually greater than the percentage falling into Organization.

In choosing reasons for their resolved scores, some committee members also inserted verbal comments. Most comments addressed problems with the examinee response that either e-rater-T or the first human rating failed to take into account. For Content, among the most frequently stated comments were "Does not fully address the prompt," "underdeveloped," "needs more development," "does not address prompt," "insufficient details to determine understanding of prompt," and "details provided are irrelevant to prompt." Also frequently cited but only with respect to the subsample of responses whose scores were resolved in favor of the first human rating were "entire essay is verbatim from article," "rewrote prompt," and "prompt regurgitation." The reasons suggest instances in which the student's response was simply a restatement of the prompt that was scored higher by e-rater-T than by the first human rating. For Organization, the frequently cited comments included "poorly organized," "poorly organized and confusing," "poor organization with severe mechanical errors that impede understanding," "list-like," "unevenly organized," and "repetitive." These comments were not associated with a particular type of resolved score.

How Well Does the Theoretically Driven Scoring Model Developed for One NAEP Prompt Generalize to Other NAEP Prompts of the Same Genre?

To address this question, the e-rater-T scoring model created for grading the informative essay prompt (Essay 1) was used for scoring two additional prompts from that genre. In addition, the e-rater-T scoring model created for grading the persuasive prompt (Essay 2) was employed for scoring two new prompts from that genre. Within each genre, the first new prompt (designated informative 1 and persuasive 1) was selected to match the characteristics of the original prompt as closely as possible. Finally, e-rater-E and e-rater-H models were used to score the responses to each of the four new prompts using the features and weights derived by those programs for evaluating the original prompts.

The generalizability of each scoring approach was evaluated by comparing the different e-rater scores to human scores obtained from the main NAEP data files for those same responses. This comparison was done for each of the four prompts separately. The indices compared included the score means, the correlations between the human scores and the e-rater scores, and the percentages exact agreement between the e-rater scores and the human scores. In principle, e-rater-T scores should be no different from, and ideally better than, the other (automated) approaches in their relations to human scores.

TABLE 27

Committee Members' Reasons for Choosing a Resolved Score as a Percentage of the Total Number of Reasons, Summed across All Members of a Committee and Responses to an Essay

		Reasons as a Percentage of Total Number				
	Total Number of Reasons	Content	Organ-ization	Word Choice	Mechan-ics	Other (e.g., style, audience)
Resolved in Favor of Human R1						
Committee 1, Essay 1	127	39	34	7	13	7
Committee 1, Essay 2	134	38	32	11	10	8
Committee 2, Essay 1	159	36	29	13	22	1
Committee 2, Essay 2	155	53	22	5	15	5
Total	**575**	**41**	**29**	**9**	**15**	**5**
Resolved in Favor of e-rater-T						
Committee 1, essay 1	79	47	31	7	9	7
Committee 1, essay 2	82	52	30	9	9	0
Committee 2, essay 1	80	56	21	5	17	1
Committee 2, essay 2	60	52	21	1	8	18
Total	**301**	**52**	**26**	**5**	**11**	**6**
Resolved in Favor of Neither						
Committee 1, essay 1	34	38	44	0	15	3
Committee 1, essay 2	24	23	29	17	23	9
Committee 2, essay 1	65	49	27	3	14	7
Committee 2, essay 2	85	47	22	5	19	7
Total	**208**	**39**	**30**	**6**	**18**	**6**

Note. Human R1 = first human rating. A separate sample of 60 responses was selected for each committee and essay. Committee 1 reviewed discrepant responses for e-rater-T1 and committee 2 reviewed discrepant responses for e-rater-T2.

Table 28 shows the summary statistics for the different methods. Differences among the mean scores were tested using a separate, repeated-measures ANOVA for each essay, with scoring method as the independent variable and score as the dependent variable. Results showed scoring method to be significant for all four essays (informative 1 $F_{4,796} = 80.3$, $p < .001$, informative 2 $F_{4,778} = 69.5.2$, $p < .001$, persuasive 1 $F_{4,792} = 78.0$, $p < .001$, persuasive 2 $F_{4,796} = 127.3$, $p < .001$).

Follow-up tests were conducted by contrasting each automated approach against the first human rating (see Appendix I). These tests showed that e-rater-E's mean scores were significantly lower than the human mean scores only for the persuasive 2 prompt. In contrast, e-rater-H's mean scores were significantly lower than the human scores for all prompts except persuasive 1. Finally, the e-rater-T mean scores were significantly lower for all four prompts. The effects observed for e-rater-T1 were consistently larger than those observed for e-rater-E and e-rater-H, though the size of these effects could in most instances be considered to be "small" (i.e., less than .5 standard deviation units). The effects found for e-rater-T2, however, were far larger and more practically important than those observed for the other automated approaches (effect size range = .58 to 1.13 standard deviation units).

TABLE 28

Summary Statistics for Human and Automated Scores for Students in the Generalization Samples

	Informative 1 (N = 200)		**Informative 2 (N = 198)**		**Persuasive 1 (N = 199)**		**Persuasive 2 (N = 200)**	
Scoring Method	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Human R1	3.8	1.0	3.8	1.1	3.5	1.1	3.4	1.0
e-rater-E	3.9	0.9	3.8	0.8	3.5	0.9	3.2	0.8
e-rater-H	3.6	1.2	3.4	1.1	3.4	1.1	3.2	1.2
e-rater-T1	3.4	1.1	3.2	1.0	3.3	1.8	3.1	1.1
e-rater-T2	2.9	1.1	3.0	1.0	2.7	1.0	2.3	0.9

Note. Human R1 = first human rating.

Table 29 gives the correlations of the automated approaches with the first human rating. (The correlations and percentages exact agreement among the automated methods themselves are given in Appendix I.) Surprisingly, the correlations did not appear to have attenuated appreciably from those observed for the original essays (see Table 19). The correlations with the original essays ranged from .59 to .67 for essay 1 (as compared with .50 to .75 in the generalization sample), and from .68 to .73 for essay 2 (as compared with .59 to .71 in the generalization sample). Also, in the generalization sample, there were generally no significant differences between e-rater-T1 and e-rater-H, or between e-rater-T1 and e-rater-E, in the strength of the relationship with human scores. In other words, across all essays and samples, e-rater-T1

was related about as highly to the first human rating as was e-rater-E or e-rater-H to that human rating. For e-rater-T2, however, the correlation with the first human rating was significantly lower for three of the four essays than was the correlation of e-rater-H with the human ratings (t range = -2.60 to -5.27, df range = 195 to 197, $p < .05$). e-rater-T2 was significantly less related to human scores than was e-rater-T1 only for the two informative essays ($t_{197} = -4.45$, $p < .01$ for informative essay 1 and $t_{195} = -4.40$, $p < .01$ for informative essay 2).

TABLE 29

Correlations of the Automated Scoring Approaches with the First Human Rating for the Generalization Samples

	e-rater-E	**e-rater-H**	**e-rater-T1**	**e-rater-T2**
Informative 1 (N = 200)	.66	.75	.73	.58
Informative 2 (N = 198)	.67	.67	.65	.50
Persuasive 1 (N = 199)	.60	.62	.59	.59
Persuasive 2 (N = 200)	.67	.71	.68	.63

The last analysis from the generalization sample is shown in Table 30. That table gives exact agreement percentages between each automated approach and the first human rating. Except for e-rater-T2, the values were not very different from the ones observed with the original essays. For essay 1, those original values ranged from 39 to 43 (as compared with 33 to 52 in the generalization sample). For essay 2, the original values were 44 to 50 (as contrasted with 39 to 53). In contrast, e-rater-T2's original percentages were 35 for essay 1 (compared to 22 and 24 in the generalization sample) and 37 for essay 2 (compared to 25 and 21 in the generalization sample).

Considering only the generalization samples, the e-rater-T1 exact-agreement percentages appeared to be lower than the exact-agreement percentages for e-rater-E and e-rater-H. e-rater-T1's agreement with the first human rating was, for three of the four essays, between 9 and 13 points lower than e-rater-H's agreement with the first rater and between 7 and 16 points lower than e-rater-E's agreement with the first rater. (For the fourth essay, there was no difference in agreement percentages vis-à-vis e-rater-H and only 3 points difference with respect to e-rater-E.) e-rater-T2's agreement was between 19 and 31 points lower than e-rater-H's agreement with the first human rating and 22 to 32 points lower than e-rater-E's agreement with the first rating.

TABLE 30

Percentage of Exact Agreement of the Automated Essay Scoring Approaches with the First Human Rating for the Generalization Samples

	e-rater-E	**e-rater-H**	**e-rater-T1**	**e-rater-T2**
Informative 1 (N = 200)	50	52	43	22
Informative 2 (N = 198)	49	46	33	24
Persuasive 1 (N = 199)	47	44	44	25
Persuasive 2 (N = 200)	53	52	39	21

Discussion

The objective of this study was to lay the groundwork for a more theoretically driven approach to automated essay scoring. The study grew out of the conviction that the defensibility of automated essay scoring is not simply a function of the ability to predict the scores that a human rater would assign but to do so for the right reasons. The practical importance of such an approach is in potentially providing a more credible and educationally meaningful method for automatically scoring writing assessments that NAEP can apply once it begins collecting essay responses in digital form.

The study evaluated a method for scoring NAEP writing assessments automatically in which weights were set by expert judgment rather than by statistical methods. This approach was compared to a brute empirical one in which both the selection of writing features and their weights were determined to be statistically optimal and to a hybrid approach in which the features were fixed but the weights were determined empirically.

Three research questions were addressed. The first question related to the extent to which judgmentally determined weights were reproducible. Two expert committees independently weighted five writing dimensions on a 0-100 scale, producing weights that were initially very similar. Further, the initial weights assigned by the two committees were much closer to one another than either committee's weights were to the hybrid approach's empirical weights. The differences between the committees' initial weights and the hybrid's empirical weights were stark: the committees believed that between 63% and 71% of the essay score should be based on Organization and development and Topical analysis. The empirical weights, in contrast, gave only 20%-21% of the emphasis to these dimensions. Instead Grammar, usage, mechanics, and style and Essay length received 69% to 73% of the empirical weight, while the committees awarded only 20% to 26% of the weight to the combination of these dimensions.

These results are consistent with two propositions. The first proposition is that expert committees have generally similar views as to what dimensions are more or less important in defining good writing for 8th grade students. The second proposition is that the views of such expert committees are not necessarily what would emerge from a more atheoretical, statistically optimal weighting of those same dimensions.

The high agreement between the two committees noted above applies to the dimension weights *initially* selected by each committee. As the weighting process proceeded, both committees received information about the way in which the dimensions were measured in the automated scoring, and one committee saw the empirical weights used by the hybrid approach for those same dimensions. Upon selecting its final weights, this committee came closer in its judgments to the empirical weights and diverged more from the other committee. Even so, the empirical weights still gave greater emphasis to Grammar, usage, mechanics, and style and to Essay length than either committee did. Similarly, the empirical weights gave less consideration to Organization and development and to Topical analysis than did either committee.

The second study question concerned how the three approaches to automated scoring compared to one another in their relations to human scores and to other indicators. Two versions of the theoretical approach were implemented, as the final weights produced by the expert committees appeared to diverge from one another enough and it was not possible to know what the impact on scores of this divergence would be. The theoretically based version derived from the committee that was aware of the hybrid's weights was dubbed e-rater-T1. The version derived by the committee independently of knowing the hybrid's weights was called e-rater-T2.

Four categories of analysis were conducted to compare the approaches. These categories concerned relations with human scores, relations with other indicators, functioning in NAEP reporting groups, and resolution of large machine-human score discrepancies. The associated analyses entailed many statistical tests, a small number of which would be expected to reach statistical significance by chance alone. Table 31 summarizes results from the four categories, but for only those analyses that showed consistent differences in functioning for e-rater-T scores *across the two essays*. (Analyses for which there were no consistent differences for either e-rater-T version are not shown in the table.) Further, the following discussion centers on the larger pattern of results across the four categories of analysis.

For e-rater-T1, there were no consistent mean score differences with human ratings; no differences in its correlations with human ratings as compared to the correlations of the other automated approaches and human ratings, and no difference between its inter-prompt correlation and the human inter-prompt correlation. e-rater-T1 also did not differ from the humans in its correlations with such indicators as typing speed, writing activities, reading activities, main NAEP writing performance, or main NAEP reading performance. Although there were some consistent differences, these were often small. For example, e-rater-T1 did not agree as highly with humans as did the hybrid or empirical versions (but the difference was only 2-6 points in percentage of exact agreement). Also, T1 functioned differentially for males, giving them mean scores that were lower than human scores (though by less than .2 standard deviation units).

In contrast to e-rater-T1, e-rater-T2 showed many consistent differences in functioning, some of which were quite substantial. e-rater-T2 produced mean scores that were significantly lower than human scores; correlated less with human scores than did the hybrid version; had considerably lower rates of exact agreement with humans than did either the brute empirical or hybrid versions; and had a lower between-prompt correlation than observed for human scores. e-rater-T2 correlated less with main NAEP reading performance than did human ratings and e-rater-T2 awarded lower scores to male examinees than human raters awarded. Among a sample of responses with large machine-human discrepancies, e-rater-T2's correlations with the resolved scores were dramatically lower than both the hybrid's and the empirical approach's correlations

TABLE 31

Consistent Differences between e-rater-T and Other Automated Approaches in Their Relations to Human Scores and Other Indicators

Analysis	e-rater-T1	e-rater-T2
Relations with Human Scores		
Mean differences		T2 < human by .16-.24 SD
Correlations with human scores		T2 < hybrid by .04-.08 points
Percentage of exact agreement with human scores	T1 < hybrid by 3-7 points T1 < empirical by 2-6 points	T2 < hybrid by 6-14 points T2 < empirical by 8-12 points
Inter-prompt correlations		T2 < human by .06 points
Relations with Other Indicators		
Correlation with Main NAEP reading performance		T2 < than human by .10 and .08 points
Correlation with Essay length	T1 > than other automated approaches T1 > than human	T2 < than other automated approaches T2 > than human
Functioning in NAEP Reporting Groups		
Mean differences	For males, T1 < human by .08 and .19 SD	For males, T2 < human by .24 and .27 SD
Large Machine-Human Score Discrepancies		
Correlations with resolved scores		T2 < hybrid by .35 and .19 points T2 < empirical by .30 and .27 points

Note. Empirical = e-rater-E. Hybrid = e-rater-H.

with the resolved scores. Finally, e-rater-T2 proved to be significantly less related to the brute empirical and hybrid versions than was e-rater-T1.

The last study question related to how well the theoretically driven scoring model developed for one NAEP prompt generalized to other NAEP prompts of the same genre. Some significant degree of generalizability across prompts in a genre should be expected if the judgmentally generated feature weights have broader theoretical meaning. Table 32 summarizes the results for the e-rater-T scores. e-rater-T1's correlations with human scores did not differ significantly from the correlations with humans of the hybrid and empirical approaches. e-rater-T1 did award mean scores that were lower than human scores by small amounts (but the hybrid approach's mean scores also were significantly lower than the human mean scores for three of these same four prompts). Finally, for three of the four prompts, T1 had percentages of exact agreement with humans that were lower by moderate amounts than both the hybrid and empirical exact agreements.

e-rater-T2 showed more and larger differences. It awarded lower mean scores than humans by moderate to large amounts. It had dramatically lower percentages of exact agreement with humans for three of four essays than the hybrid approach's agreement or the brute empirical approach's agreement with human ratings. Finally, with respect to the rank ordering of scores, e-rater-T2 correlated significantly lower with the human ratings for three of four essays than did the hybrid with that same human rating.

TABLE 32

Consistent Differences between e-rater-T and Other Automated Approaches in the Generalization Samples

Analysis	**e-rater-T1**	**e-rater-T2**
Mean differences	T1 < human by .21-.53 SD	T2 < human by .58-1.13 SD
Correlations with human scores		For three of four essays, T2 < hybrid by .08-.18 points
Percentage of exact agreement with human scores	For three of four essays, T1 < hybrid by 9-13 points	For all four essays, T2 < hybrid by 19-31 points
	For three of four essays, T1 < empirical by 7-16 points	For all four essays, T2 < empirical by 22-32 points

Note. Empirical = e-rater-E. Hybrid = e-rater-H.

Overall, then, the results seem to suggest that the two versions of the theoretical approach operated differently from one another. e-rater-T1, based on the judgments of a committee that had access to the hybrid weights, produced scores that showed relatively few consistent differences from the hybrid approach (or from the brute empirical one), at least on the two original essays. The lack of consistent differences is probably because the committee chose weights that were similar to those used by e-rater-H. (The correlations between the e-rater-T1 and e-rater-H scores were in the low .90s for both of the two original essays.) The e-rater-T1

scores were, however, somewhat less generalizable than the ones coming from the hybrid and from the brute empirical approaches. This result suggests that statistically optimal weights (and features) may remain more stable across prompts, examinees, and raters than judgmentally derived weights.

That statistically optimal weights retain their stability is not necessarily testament to their theoretical meaningfulness. For example, this result may mean nothing more than that operational conditions cause human raters to attend to the same features in the same proportions from one prompt to the next. Grammar, usage, mechanics, and style errors, which e-rater-H weighted highly in this data set, may be one such collection of features. In operational grading, a premium is placed on speed and on agreement among raters. Errors like these are an attractive focus for raters because they are easily, quickly, and objectively detectable.

Thus, it may be the case that empirical weights can provide a useful starting point for expert committees, with the understanding that the committee would moderate the weights only somewhat to bring them more into line with theoretical considerations. Under such circumstances, the results may turn out to be reasonable in the sense of being both more acceptable to writing experts and not too divergent from what an operational scoring would normally produce.

Of course, an intended gain in theoretical meaningfulness may not occur if the manner in which the automated scoring implements its dimensions is only superficially consistent with theory. And, in fact, our expert committees raised a number of questions about the completeness of e-rater 2.1's coverage, in particular the very limited attention to style, the view of organization in terms of the five-paragraph model, and the neglect of audience awareness.

Further, results may look less positive than they otherwise might if the operational scoring rubric itself is in some way lacking and human readers faithfully follow that rubric. Indeed, our committee members commented about problems they perceived with the NAEP rubrics. These problems included that the criteria for scoring informative and persuasive essays differed only marginally, the informative rubric did not include quality of ideas or content, the persuasive rubric did not credit for acknowledging another point of view, appropriateness for the intended audience was not considered, and the performance standards seemed too low.

Finally, we should not be deceived into thinking that human and automated scores mean the same thing. Human and automated scores differ often enough in exact agreement and in rank order that they could be measuring somewhat different constructs, as the results of this study suggest. As one example, *all* of the automated methods correlated notably higher with essay length than did the human ratings. As a second example, the correlation of the brute empirical approach with main NAEP writing performance, arguably the most credible indicator of writing skill employed in this study, was significantly lower than the correlation of human scores with NAEP performance. Last, the experts' resolutions of large machine-human score discrepancies usually correlated higher with the human ratings than with the automated scores, and the most common reasons for these resolutions were issues of content and organization.

What are the implications of this study for NAEP? To provide an accurate representation of how effectively the nation's students write, NAEP will inevitably need to include measures of writing on computer (Horkay et al., 2005). At that time, it will become possible to score results automatically, which could decrease costs and reporting cycles substantially. That scoring can be arranged to predict optimally the judgments that human raters would assign. This study

suggests, however, that it is possible to adjust the parameters of automated scoring to bring them at least somewhat more into line with the values of writing experts and still produce credible results. Such adjustments essentially constitute a construct redefinition. That is to say that the construct measured by a NAEP writing assessment is not necessarily the one the rubric describes but the one that NAEP readers implement. Automated scoring with parameters adjusted by writing experts may allow that construct definition to be more precisely described, more openly debated, and more carefully implemented than is the case with human rating.[7]

Future research might focus on at least two directions. One direction might be to use current theories of writing cognition to create a coherent, principled basis for deriving scoring dimensions and features. The work of Hayes and colleagues (Hayes, 1996; Hayes & Flower, 1980) represents one well-articulated theory with which to begin. A second direction is to validate scoring based on such an analysis in a multifaceted manner that, among other things, includes (1) a comprehensive expert analysis of the extent to which the features as implemented adequately cover the dimensions derived from the theory and (2) an evaluation of the relations of automated feature scores to human ratings of the same features. Such a validation serves to recast the criterion, giving less credence to holistic ratings based on a loosely described rubric and more importance to verifying that the theory itself has been implemented faithfully in the automated scoring.

Several limitations of this study should be noted. First, it used only two expert committees. Additional committees would have provided for a more credible test of the reproducibility of weights. Second, the study employed different versions of the same automated essay scoring program, e-rater, to represent three general approaches to scoring: brute empirical, hybrid, and theoretical. While using different versions of the same program afforded some control over the characteristics that could be varied, it is not clear whether a different automated scoring program would have produced similar results. In particular, some committee members did not find e-rater v2.1's implementation of its dimensions and features in keeping with their preferences, posing a classic "avoidance-avoidance" conflict. As a result, these members occasionally assigned higher weights to less inappropriate features as a means of reducing the impact on scores of the most distasteful ones. A fourth limitation is that the three automated approaches were scaled in somewhat different ways, which may account for some of the differences observed between e-rater-T and the other two approaches (see Appendices E and F). The two versions of e-rater-T, however, were scaled in exactly the same way, so the differences in functioning between them should be unaffected by this variation in scaling parameters. Finally, only three essays per genre were evaluated and at only one grade level, restricting the degree to which results can be generalized to other essays and other grades.

[7] Y. Attali (personal communication, December 1, 2005) has created an easy-to-use tool for making such adjustments to scoring models and immediately seeing their impact on score distributions.

References

Allen, N., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater version 2.0* (ETS RR-04-45). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing*. Hillsdale, NJ: Erlbaum.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction* (ETS RR-98-15). Princeton, NJ: Educational Testing Service.

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing.* Retrieved December 16, 2005, from http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf

Burstein, J., Chodorow, M., & Leacock, C. (2004, Fall). Automated essay evaluation: the Criterion Online writing service. *AI Magazine*. Retrieved September 26, 2005, from http://www.findarticles.com/p/articles/mi_m2483/is_3_25/ai_n6258424

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal*, *93*(4), 47-52.

Cizek, G. J., & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English, 24,* 67-87.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

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

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

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

E-rater [Computer software]. (1997). Princeton, NJ: Educational Testing Service.

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

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments, and Computers, 28*(2), 197-202.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). Retrieved December 16, 2005, from http://imej.wfu.edu/articles/1999/2/04/

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy. & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum.

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan et al. (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005-457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved October 25, 2005, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Intelligent Essay Assessor [Computer software]. (1997). Boulder, CO: University of Colorado.

IntelliMetric Engineer [Computer software]. (1997). Yardley, PA: Vantage Technologies.

Johnson, R. L., Penny, J., & Gordon, B. (1999, April). *Score resolution and score reliability: An empirical study of an analytic scoring rubric*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

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

Landauer, T. K., Laham, D., & Foltz, P. W. (2001). *The Intelligent Essay Assessor: Putting knowledge to the test*. Paper presented at the Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ.

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

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic

Analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

National Commission on Writing in America's Schools and Colleges. (2003). *The neglected "R": The need for a writing revolution.* New York: College Entrance Examination Board. Retrieved April 21, 2004, from http://www.writingcommission.org/prod_downloads/writingcom/neglectedr.pdf

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

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

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Erlbaum.

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) *Automated scoring of written essays: Can such scores be valid?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002) Stumping E-Rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18,* 103-134.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait rating for automated essay scoring. *Educational and Psychological Measurement, 62,* 5-18.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment and Evaluation in Higher Education, 26(3),* 247-259.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Appendix A

Essay Prompts and Scoring Rubrics

Informative Essay ("Save a Book")

A novel written in the 1950's describes a world where people are not allowed to read books. A small group of people who want to save books memorize them so that the books won't be forgotten. For example, an old man who has memorized the novel *The Call of the Wild* helps a young boy memorize it by reciting the story to him. In this way, the book is saved for the future.

If you were told that you could save just one book for future generations, which book would you choose?

Write an essay in which you discuss which book you would choose to save for future generations and what it is about the book that makes it important to save. Be sure to discuss in detail why the book is important to you and why it would be important to future generations.

Informative Scoring Guide
Score & Description

Excellent-6

- Develops and shapes information with well-chosen details across the response.
- Well organized with strong transitions.
- Sustains variety in sentence structure and exhibits good word choice.
- Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.

Skillful-5

- Develops and shapes information with details in parts of the response.
- Clearly organized, but may lack some transitions and/or have occasional lapses in continuity.
- Exhibits some variety in sentence structure and some good word choices.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Sufficient-4

- Develops information with some details.
- Organized with ideas that are generally related, but has few or no transitions.
- Exhibits control over sentence boundaries and sentence structure, but sentences and word choice may be simple and unvaried.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Uneven-3

May be characterized by one or more of the following:

- Presents some clear information, but is list-like, undeveloped, or repetitive OR offers no more than a well-written beginning.
- Unevenly organized; the response may be disjointed.
- Exhibits uneven control over sentence boundaries and sentence structure; may have some inaccurate word choices.
- Errors in grammar, spelling, and punctuation sometimes interfere with understanding.

Insufficient-2

May be characterized by one or more of the following:

- Presents fragmented information OR may be very repetitive OR may be very undeveloped.
- Very disorganized; thoughts are tenuously connected OR the response is too brief to detect organization.
- Minimal control over sentence boundaries and sentence structure; word choice may often be inaccurate.
- Errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation interfere with understanding in much of the response.

Unsatisfactory-1

May be characterized by one or more of the following:

- Attempts to respond to prompt, but provides little or no coherent information; may only paraphrase the prompt.
- Has no apparent organization OR consists of a single statement.
- Minimal or no control over sentence boundaries and sentence structure; word choice may be inaccurate in much or all of the response.
- A multiplicity of errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation severely impedes understanding across the response.

Persuasive Essay ("School Schedule")

Imagine that the article shown below appeared in your local newspaper. Read the article carefully, then write a letter to your principal arguing for or against the proposition that classes at your school should begin and end much later in the day. Be sure to give detailed reasons to support your argument and make it convincing.

Studies Show Students Need To Sleep Late

Night Owls Versus Early Birds

The *Journal of Medicine* announced today the results of several recent studies on the sleep patterns of teenagers and adults. These studies show that adults and teenagers often have different kinds of sleep patterns because they are at different stages in the human growth cycle.

The study on teenagers' sleep patterns showed that changes in teenagers' growth hormones are related to sleeping patterns. In general, teenagers' energy levels are at their lowest in the morning, between 9 a.m. and 12 noon. To make the most of students' attention span and ability to learn, the study showed that most teenagers need to stay up late at night and to sleep late in the morning. They called this pattern "the night owl syndrome."

Studies of adults (over 30 years of age) showed the opposite sleep pattern. On average, adults' energy levels were at their lowest at night between 9 p.m. and 12 midnight and at their highest between 6 and 9 a.m. In addition, a study of adults of different ages revealed that as adults get older they seem to wake up earlier in the morning. Thus, adults need to go to sleep earlier in the evening. Researchers called this sleep pattern "the early bird syndrome."

Researchers claim that these studies should be reviewed by all school systems and appropriate changes should be made to the daily school schedule.

Persuasive Scoring Guide
Score & Description

Excellent-6

- Takes a clear position and develops it consistently with well-chosen reasons and/or examples across the response.
- Well organized with strong transitions.
- Sustains variety in sentence structure and exhibits good word choice.
- Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.

Skillful-5

- Takes a clear position and develops it with reasons and/or examples in parts of the response.
- Clearly organized, but may lack some transitions and/or have occasional lapses in continuity.
- Exhibits some variety in sentence structure and some good word choices.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Sufficient-4

- Takes a clear position and supports it with some reasons and/or examples.
- Organized with ideas that are generally related, but there are few or no transitions.
- Exhibits control over sentence boundaries and sentence structure, but sentences and word choice may be simple and unvaried.
- Errors in grammar, spelling, and punctuation do not interfere with understanding.

Uneven-3

May be characterized by one or more of the following:

- Takes a position and offers support, but may be unclear, repetitive, list-like, or undeveloped.
- Unevenly organized; the response may be disjointed.
- Exhibits uneven control over sentence boundaries and sentence structure; may have some inaccurate word choices.
- Errors in grammar, spelling, and punctuation sometimes interfere with understanding.

Insufficient-2

May be characterized by one or more of the following:

- Takes a position, but may be very unclear, very undeveloped, or very repetitive.
- Very disorganized; thoughts are tenuously connected OR the response is too brief to detect organization.
- Minimal control over sentence boundaries and sentence structure; word choice may often be inaccurate.
- Errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation interfere with understanding in much of the response.

Unsatisfactory-1

May be characterized by one or more of the following:

- Attempts to take a position (addresses topic) but is incoherent OR takes a position but provides no support; may only paraphrase the prompt.
- Has no apparent organization OR consists of a single statement.
- Minimal or no control over sentence boundaries and sentence structure; word choice may be inaccurate in much or all of the response.
- A multiplicity of errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation severely impedes understanding across the response.

Appendix B

e-rater Dimensions and Feature Descriptions

Dimension	Feature	Description
Grammar, usage, mechanics, & style	Ratio of grammar errors to total words in the essay	This feature is a transformation of the rate of sentence fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, possessive errors, and wrong or missing words (e.g., "the" instead of "they") in the essay.
	Ratio of mechanics errors to total words in the essay	This feature is a transformation of the rate of errors in spelling, capitalization, punctuation, hyphenation, duplicate words (e.g., "the the), and compound words in the essay.
	Ratio of usage errors to total words in the essay	This feature is a transformation of the rate of wrong articles, missing or extra articles, confused words (e.g., "there" and "their"), wrong word forms, faulty comparisons (e.g., "most rarest"), preposition errors, and nonstandard word forms (e.g., "gonna") in the essay.
	Ratio of style errors to total words in the essay	This feature is a transformation of the rate of repetitious words, too many sentences beginning with coordinating conjunctions, too many short sentences, too many long sentences, and use of passive voice in the essay.
Organization & development	The number of "discourse" units out of 8	This feature indicates how many discourse units the essay has relative to an "optimal" number. The "optimal" number includes a *thesis* statement, three *main idea* units, a *supporting idea* unit for each main idea unit, and a *conclusion*.
	The average length of the discourse units	This feature indicates, on average, how long is the discussion that comprises each discourse unit. The feature is the total number of words in the essay divided by the number of discourse units.
Topical analysis	Similarity of the essay's content to other previously scored essays in the top score category (6)	This feature indicates how similar the essay's vocabulary is to the vocabulary of the best essays (where which essays are considered best is based on the scores of human judges).
	The score category (1-6) containing essays whose words are most similar to the target essay	This feature indicates the score level to which the essay's text is most similar with regard to vocabulary.

Dimension	Feature	Description
Word complexity	Word repetition	This feature is computed by dividing the total number of different words in an essay by the total number of all words. It is equivalent to the type/token ratio.
	Vocabulary difficulty	This feature is based on word *infrequency*, the supposition being that, all other things equal, the use of infrequent words indicates more sophisticated vocabulary. The frequency index is a corpus-based measure that employs the Lexile index. Using the Lexile index, all of the words in the essay are assigned a frequency value. The value for the word with the fifth lowest value is used to represent the essay's vocabulary difficulty.
	Average word length (computed across all words in the essay)	
Essay length	Total number of words in the essay	

Appendix C

Expert Committee Members

Marcia Ashhurst-Whiting, New Jersey Department of Education

Anthony Bucco, Paramus (NJ) Schools

Gail Hawisher, University of Illinois (Champaign-Urbana)

Geof Hewitt, Vermont Department of Education

Brian Huot, University of Louisville

Tanji Reed Marshall, Charlotte-Mecklenburg (NC) Schools

Brian Medley, Camden (NJ) Schools

Patricia McGonegal, Mt. Mansfield (VT) High School

Lee Odell, Rensselear Polytechnic Institute

Appendix D

Transforming e-rater Scores to the 1-6 Human-Rater Scale

e-rater-E

The standard procedure for scaling e-rater v1.3 scores was employed to place scores on the human-rater scale. The procedure was as follows:

1. For each response in the training sample, compute the mean of the human ratings, where more than one rater scored the response. If only one rater scored the response, use only that rater's score in place of the mean human score.
2. Use stepwise linear regression to produce an equation containing that subset of e-rater features most predictive of the mean human scores.
3. In the cross-validation sample, use e-rater v1.3 to produce feature scores for each response.
4. For each response, enter the feature scores into the regression equation to produce a continuous total score on the 1-6 human-rater scale.
5. Round each continuous total score using the following default cut-points: 1.5, 2.5, 3.5, 4.5, 5.5.

e-rater-H

The procedure most frequently used in operational scaling to date was employed for placing e-rater v2 scores on the human-rater scale. The procedure was as follows:

1. For each essay in the training sample, compute the mean of the human ratings where more than one rater scored the essay. If only one rater scored the response, use only that rater's score in place of the mean human score.
2. Use linear regression to produce an equation that weights the 11 e-rater v2.1 features to best predict the mean human scores. (Fix the weight for the 12^{th} feature, length, to 30%, a commonly used operational default.)
3. Produce relative weights for all 12 features as per Attali and Burstein (2005).
4. In the cross-validation sample, use e-rater v2.1 to produce feature scores for each essay.
5. For each essay, enter the standardized feature scores into the regression equation to produce a continuous total score.
6. Rescale the resulting distribution of continuous scores to the 1-6 scale using the mean and standard deviation of the mean human scores in the training sample.
7. Round the continuous rescaled e-rater scores using cut points determined for each prompt and examinee sample by an algorithm created by Y. Attali. For essay 1, the cut points (rounded to one decimal place) were: 2.1, 2.5, 3.3, 4.4, 5.6. For essay 2, they were: 2.1, 2.3, 3.3, 4.5, 5.4.

e-rater-T

For each variation of e-rater-T, the transformation for each of the two WOL essay prompts in turn was computed using the following steps:

1. Standardize each feature score in the training sample to a mean of 0 and a standard deviation of 1.
2. For each response, multiply the appropriate weight set by the relevant study committee by each feature score, where the weight for a feature is the mean weight taken across committee members multiplied by the mean dimension weight taken across committee members.
3. For each response, sum the weighted feature scores to produce a continuous total score for that response.
4. Rescale the resulting distribution of continuous total scores to the 1-6 scale employed by NAEP using the mean and standard deviation of the first human rating for the training sample responses.
5. Standardize each feature score in the cross-validation sample using the mean and standard deviation previously calculated for that feature in the training sample.
6. Multiply the appropriate committee weight by each standardized feature score and sum the weighted standardized feature scores to form a continuous total score.
7. Rescale the continuous total score using the scaling parameters as determined from the training sample in step 4.
8. Round the rescaled continuous total scores using cut points determined through Y. Attali's approximation to the algorithm used operationally for e-rater v.2.1. For both essays, those cut-points (rounded to one decimal place) were: 1.7, 2.6, 3.5, 4.5, 5.3.

Appendix E

Impact of Scaling Differences on e-rater-T Scores

The procedure used for scaling e-rater-T scores differed from that used to scale e-rater-E and e-rater-H scores in two major ways. First, both e-rater-E and e-rater-H were scaled to the mean of two human ratings (where there were two human ratings). This procedure is common operational scaling practice for these systems, which were originally deployed in settings where two human ratings were used operationally (e.g., GMAT). e-rater-T scores, on the other hand, were scaled to the first human rating, which is the only rating used by NAEP for operational scoring purposes. (The second rating is employed only for estimating interrater reliability.) In the training sample, which was used to provide the scaling parameters, 55 of 250 responses had a second rating for essay 1 and 67 of 250 responses had a second rating for essay 2. As Table E-1 shows, the differences in distributions appear to be minimal.

TABLE E-1

Summary Statistics for Different Configurations of Human Rater Scores in the Training Sample (N = 250)

Human Score	**Mean**	**SD**
Essay 1		
Human R1	3.55	1.35
Human R1 + R2	3.56	1.33
Essay 2		
Human R1	3.44	1.31
Human R1 + R2	3.46	1.30

Note. Human R1 = first human rating. Human R1 + R2 = the mean of two human ratings, where two human raters scored an essay; otherwise the first human rating is used.

The second major difference between the procedure used for scaling e-rater-T scores and those used to scale e-rater-E and e-rater-H scores was in the cut points, which were particular to each approach (see Appendix D).

Table E-2 shows the combined impact on e-rater-T scores in the cross-validation sample of both scaling to the mean of the two human scores (from the training sample) and of applying the e-rater-E and e-rater-H cut points. The first row in the table under each essay gives the summary statistics for scores as calculated for this study. As the table shows, the *largest* difference in means between the procedure used in the study and either alternative scaling is .08 points for committee 1 on essay 2. The largest difference in standard deviations is .09 points, which occurs for both committees on essay 2.

TABLE E-2

Summary Statistics in the Cross-Validation Sample for e-rater-T Scores Scaled in Three Ways (N = 1,005)

Scaling Method	**Committee 1**		**Committee 2**	
Essay 1	**Mean**	**SD**	**Mean**	**SD**
Human R1 with e-rater-T cut-points	3.59	1.26	3.41	1.23
Human R1 + R2 with e-rater-E cut-points	3.58	1.18	3.42	1.16
Human R1 + R2 with e-rater-H cut-points	3.61	1.28	3.46	1.29
Essay 2				
Human R1 with e-rater-T cut-points	3.32	1.23	3.22	1.27
Human R1 + R2 with e-rater-E cut-points	3.35	1.14	3.24	1.18
Human R1 + R2 with e-rater-H cut-points	3.40	1.24	3.25	1.30

Note. Human R1 = first human rating. Human R1 + R2 = the mean of the two human ratings, where two human raters scored an essay; otherwise the first human rating is used.

Finally, Table E-3 gives the correlations between the e-rater-T scores produced by the procedure used in the study and the two alternative scalings. As can be seen, there are only small differences in the rank orders that the scalings produce.

TABLE E-3

Correlations between e-rater-T Scores as Scaled in the Study and e-rater-T Scores Scaled by Two Alternative Procedures in the Cross-Validation Sample (N = 1,005)

Scaling Method		
Essay 1	**Committee 1**	**Committee 2**
Human R1 + R2 with e-rater-E cut-points	.98	.98
Human R1 + R2 with e-rater-H cut-points	.95	.94
Essay 2		
Human R1 + R2 with e-rater-E cut-points	.98	.98
Human R1 + R2 with e-rater-H cut-points	.93	.94

Note. Human R1 + R2 = the mean of the two human ratings, where two human raters scored an essay; otherwise the first human rating is used.

Appendix F

Post-Hoc Contrasts for Automated Methods against the First Human Rating

TABLE F-1

Post-Hoc Contrasts for Difference between the Mean Score Assigned by Each Automated Method and the First Human Rating in the Total Cross-Validation Sample (N = 1,005)

	Essay 1		Essay 2	
Contrast	$F_{1,1004}$	**Effect Size**	$F_{1,1004}$	**Effect Size**
e-rater-E vs. human R1	0.7	--	11.6*	.09
e-rater-H vs. human R1	5.7*	-.06	7.5*	.06
e-rater-T1 vs. human R1	0.4	--	46.5*	.16
e-rater-T2 vs. human R1	32.1*	.16	89.0*	.24

*$p < .05$.

Note. Effect sizes are in standard deviation units and are given only for significant effects. Human R1 = first human rating.

Appendix G

Summary Statistics for Other Indicators

TABLE G-1

Summary Statistics for Indicators Used to Distinguish the Functioning of the Different Automated Scoring Approaches in the Cross-Validation Sample

Indicator	**N**	**Scale**	**Mean**	**SD**
Typing speed	948	0-78	39	18
Writing activities	629	0-1	.55	.14
Reading activities	279	0-1	.43	.15
Writing plausible value 1	687	0-300	157	32
Writing plausible value 2	687	0-300	156	34
Writing plausible value 3	687	0-300	157	33
Writing plausible value 4	687	0-300	155	34
Writing plausible value 5	687	0-300	157	33
Reading plausible value 1	318	0-500	267	29
Reading plausible value 2	318	0-500	267	29
Reading plausible value 3	318	0-500	270	30
Reading plausible value 4	318	0-500	267	29
Reading plausible value 5	318	0-500	267	29
Essay 1 length (in words)	1,005	0-712[a]	186	101
Essay 2 length (in words)	1,005	0-720[a]	160	86

[a] The maximums for essay length are the values for the longest essays observed.

Appendix H

Post-Hoc Contrasts for Automated Methods against the First Human Rating in NAEP Reporting Groups

TABLE H-1

Post-Hoc Contrasts of Mean Score for Each Automated Approach against the Mean of the First Human Rating, by Race/Ethnicity in the Cross-Validation Sample (N = 1005)

	Essay 1						**Essay 2**					
	White		**Black**		**Hispanic**		**White**		**Black**		**Hispanic**	
Contrast	F	Effect Size	F	Effect Size	F	Effect Size	F	Effect Size	F	Effect Size	F	Effect Size
e-rater-E vs. human R1	.2	--	15.9*	<-.26	.1	--	32.3*	>.18	.1	--	4.3*	<-.15
e-rater-H vs. human R1	.3	--	4.3*	<-.14	.1	--	27.6*	>.15	.1	--	.1	--
e-rater-T1 vs. human R1	.1	--	.2	--	.2	--	68.3*	>.23	.4	--	.7	--
e-rater-T2 vs. human R1	35.5*	>.21	.2		.8	--	85.1*	>.29	8.3*	>.17	.1	--

*$p < .05$

Note. Effect sizes are in standard deviation units and are given only for significant effects. Human R1 = first human rating.

> = the human R1 mean is higher than the e-rater mean.

< = the human R1 mean is lower than the e-rater mean.

TABLE H-2

Post-Hoc Contrasts of Mean Score for Each Automated Approach against the Mean of the First Human Rating, by Gender in the Cross-Validation Sample (N = 1005)

	Essay 1				Essay 2			
	Male		Female		Male		Female	
Contrast	**F**	**Effect Size**	**F**	**Effect Size**	**F**	**Effect Size**	**F**	**Effect Size**
e-rater-E vs. human R1	.3	--	.9	--	.4	--	16.4*	>.16
e-rater-H vs. human R1	.8	--	9.9*	<-.12	7.0*	>.08	.2	--
e-rater-T1 vs. human R1	5.5*	>.08	.2	--	37.5*	>.19	12.3*	>.13
e-rater-T2 vs. human R1	40.6*	>.24	.1	--	60.9*	>.27	29.8*	>.21

*$p < .05$

Note. Effect sizes are in standard deviation units and are given only for significant effects. Human R1 = first human rating.

> = the human R1 mean is higher than the e-rater mean.

< = the human R1 mean is lower than the e-rater mean.

TABLE H-3

Post-Hoc Contrasts of Mean Score for Each Automated Approach against the Mean of the First Human Rating by Eligibility for Free or Reduced-Price School Lunch in the Cross-Validation Sample (N = 1,005)

	Essay 1				Essay 2			
	Eligible		Not Eligible		Eligible		Not Eligible	
Contrast	**F**	**Effect Size**	**F**	**Effect Size**	**F**	**Effect Size**	**F**	**Effect Size**
e-rater-E vs. human R1	15.3*	< -.19	4.7*	> .08	.6	--	23.6*	> .16
e-rater-H vs. human R1	10.9*	< -.16	.9	--	.7	--	16.4*	> .12
e-rater-T1 vs. human R1	6.0*	< -.12	8.1*	> .10	.1	--	48.7*	> .21
e-rater-T2 vs. human R1	.9	--	48.3*	> .26	14.8*	> .19	59.1*	> .25

*$p < .05$

Note. Effect sizes are in standard deviation units and are given only for significant effects. Human R1 = first human rating.

> = the human R1 mean is higher than the e-rater mean.

< = the human R1 mean is lower than the e-rater mean.

Appendix I

Intercorrelations and Percentages of Exact Agreement Among the Automated Approaches in the Generalization Samples

Table I-1 shows the intercorrelations among the automated approaches. Two observations may be worth noting. First, the correlational pattern was strikingly similar across all four essays and samples. The singular exception was for the correlation between e-rater-E and e-rater-T2, which was in the high .50s for the two informative essays but in the high .70s for the two persuasive prompts (z = 19.01 for the difference between the informative 1 and persuasive 1 correlations and z = 19.95 for the difference between the informative 2 and persuasive 2 correlations). Second, the overall pattern was quite similar to the pattern found for the two essays on which the scoring models were originally created (see Table 17).

TABLE I-1

Intercorrelations among the Automated Essay Scoring Approaches for the Generalization Samples

	e-rater-E	**e-rater-H**	**e-rater-T1**
Informative 1 (N = 200)			
e-rater-H	.73	-	-
e-rater-T1	.76	.89	-
e-rater-T2	.59	.75	.78
Informative 2 (N = 198)			
e-rater-H	.74	-	-
e-rater-T1	.73	.89	-
e-rater-T2	.58	.76	.80
Persuasive 1 (N = 199)			
e-rater-H	.73	-	-
e-rater-T1	.77	.89	-
e-rater-T2	.70	.79	.84
Persuasive 2 (N = 200)			
e-rater-H	.73	-	-
e-rater-T1	.73	.87	-
e-rater-T2	.70	.79	.84

Table I-2 gives the percentages of exact agreement among the automated approaches. Of note in this table is that the percentages were especially variable across essays and samples for e-rater-T2. Also, in most cases, e-rater-T2's exact agreement with the other approaches appeared generally lower and sometimes more variable than it did for the original two essays (see Table 18). For those original essays, e-rater-T2's exact agreement ranged from the low 40s to high 50s for essay 1 (as compared with the low 20s to 60 in the generalization sample) and from the mid-40s to high 60s for essay 2 (as compared with the low 20s to mid 30s in the generalization sample). Because e-rater-T2's correlations with the other approaches did not appear to change dramatically from the pattern observed for the original essays, the changes observed for exact agreement suggest that its scaling of scores did not remain consistent relative to the other approaches for the generalization sample.

TABLE I-2

Percentage of Exact Agreement among the Automated Scoring Approaches for the Generalization Sample

	e-rater-E	**e-rater-H**	**e-rater-T1**
Informative 1 (N = 200)			
e-rater-H	53	-	-
e-rater-T1	46	66	-
e-rater-T2	21	31	46
Informative 2 (N = 198)			
e-rater-H	51	-	-
e-rater-T1	41	64	-
e-rater-T2	27	47	60
Persuasive 1 (N = 199)			
e-rater-H	54	-	-
e-rater-T1	58	73	-
e-rater-T2	28	32	36
Persuasive 2 (N = 200)			
e-rater-H	51	-	-
e-rater-T1	49	63	-
e-rater-T2	23	23	30

Appendix J

Post-Hoc Contrasts for Automated Methods against
the First Human Rating in the Generalization Samples

TABLE J-1

Post-Hoc Contrasts of Mean Scores for Each Automated Approach against the Mean of the First Human Rating

	Informative 1		**Informative 2**		**Persuasive 1**		**Persuasive 2**	
Contrast	**F**	**Effect Size**	**F**	**Effect Size**	**F**	**Effect Size**	**F**	**Effect Size**
e-rater-E vs. human R1	0.2	--	0.03	--	0.6	--	7.0*	.16
e-rater-H vs. human R1	9.1*	.15	34.9*	.34	2.5	--	7.2*	.15
e-rater-T1 vs. human R1	50.5*	.37	78.9*	.53	10.8*	.21	24.6*	.28
e-rater-T2 vs. human R1	149.4*	.79	99.9*	.71	157.9*	.58	335.7*	1.13

*$p < .01$

Note. Human R1 = first human rating. Effect sizes are in standard deviation units and are given only for significant contrasts.