J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 6, Number 1 · August 2007

Toward More Substantively Meaningful Automated Essay Scoring

Anat Ben-Simon & Randy Elliot Bennett

www.jtla.org

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

J·T·L·A

Volume 6, Number 1

Toward More Substantively Meaningful Automated Essay Scoring

Anat Ben-Simon & Randy Elliot Bennett

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

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

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

Preferred citation:

Ben-Simon, A. & Bennett, R.E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 6*(1). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This study evaluated a "substantively driven" method for scoring NAEP writing assessments automatically. The 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 *substantively driven* approach in which a fixed set of variables was weighted according to the judgments of two independent committees of writing experts. The research questions concerned (a) the reproducibility of weights across writing experts, (b) the comparison of scores generated by the three automated approaches, and (c) the extent to which models developed for scoring one NAEP prompt generalize to other NAEP prompts of the same genre. Data came from the 2002 NAEP Writing Online study and from the main NAEP 2002 writing assessment.

Results showed that, in carrying out the substantively driven approach, experts initially assigned weights to writing dimensions that were highly similar across committees but that diverged from one another after committee 1 was shown the empirical weights for possible use in its judgments and committee 2 was not shown those weights. The substantively driven approach based on the judgments of committee 1 generally did not operate in a markedly different way from the brute empirical or hybrid approaches in most of the analyses conducted. In contrast, many consistent differences with those approaches were observed for the substantively driven approach based on the judgments of committee 2. This study suggests that 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 substantive 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.

J·T·L·A

Toward More Substantively Meaningful Automated Essay Scoring

Anat Ben-Simon
National Institute for Testing and Evaluation
Jerusalem, Israel
Randy Elliot Bennett
Educational Testing Service
Princeton, NJ

Introduction

The National Assessment of Educational Progress (NAEP) spends significant time and monetary resources for scoring essay responses. In 2011, the NAEP writing assessment will be delivered on computer for the first time (Olson, 2007). If NAEP essays can 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 some of the 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 substantively 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 (AES)

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: (a) the type of lower-level features used by the system and, in particular, their relationship to writing characteristics grounded in a theoretical model; (b) the grouping of these features into higher-level writing dimensions; and (c) 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 may not necessarily have any direct, explicit 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 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 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 (Shermis, Koch, Page, Keith, & Harrington, 2002). This innovation was introduced to provide more detailed feedback about students' strengths and weaknesses. 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 (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 are said to fall into discourse/rhetorical, 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 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 through 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 (Landauer, Laham & Foltz, 2003):

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 grounded in a well-described theory of knowledge representation (Landauer & Dumais, 1997), and content is arguably a key factor in evaluating writing quality.

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).[1] 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 (next page), 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 counter-intuitive weights, which may be set to zero, and for essay length, which has often been 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.

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., thesis, main ideas, supporting ideas, conclusion)
	6. The average length of each unit in words
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)

Validity Issues in AES

Yang, Buckendahl, Juszkiewicz, and Bhola (2002) classify validation approaches for automated scoring into three categories: (a) approaches focusing on the relationship among scores generated by different scorers (human and computer), (b) approaches focusing on the relationship between test scores and external measures of writing, and (c) 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. (Table 2 on the next page, shows representative results.)

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).

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 pairing 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, Koch, Page, Keith, & Harrington, 2002	English placement test (1 prompt)	386	.71 (*r* between a single rating and 5 pooled ratings)	.83 (*r* between computer rating and 5 pooled ratings)
IntelliMetric	Elliot, 2001	K – 12 norm-referenced test	102	.84	.78 – .85
IEA	Landauer, Laham, & Foltz, 2003	GMAT® (1 prompt) GMAT (1 prompt)	292 285	.86 .88	.84 (1 rater) .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.

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 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 degree 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). So-called "generic" models (i.e., models generated by combining the training data from multiple prompts) would be more philosophically in line with this practice. Whereas such models have been used experimentally (e.g., Attali & Burstein, 2006), most operational testing programs still appear to use prompt-specific models.

Study Objective

The objective of this study is to lay the groundwork for a more substantively driven approach to AES that, in Yang et al.'s (2002) terms, is concerned with scoring process as well as with the empirical relations of scores. The practical importance of this substantively 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 substantively 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 an alternative approach is to be preferred on substantive grounds.
3. How well does the substantively 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 substantive meaning.

Method

Participants

The primary data set came from the NAEP Writing Online (WOL) study (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). WOL study data were collected in spring 2002 and included 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.

A secondary source of data was the eighth-grade main NAEP 2002 writing assessment. From this latter data set, approximately 300 paper 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 substantively driven models created for automatically scoring the two WOL prompts.

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.[2] 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 performance information for one nationally representative subset of the sample (N = 687 students) and main NAEP 2002 reading performance information for the other, non-overlapping, nationally representative subset (N = 568 students). The writing performance information was *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.

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

Automated Essay Scoring Approaches

Three automated scoring approaches were implemented using e-rater v1.3 and v2.1.[3] e-rater v1.3 was used to represent a brute-empirical approach (denoted as "e-rater-E"). Two configurations of e-rater v2.1 were used to represent the hybrid and substantively driven approaches and are denoted as "e-rater-H" and "e-rater-S," respectively. Table 3 summarizes the three approaches.

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. 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 substantively driven approach allows weights and, to a lesser extent, features to be determined through the judgments of writing experts.

Table 3: Three Scoring Approaches as Operationalized by Two Different Versions of e-rater

Scoring Approach	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.
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.
Substantively driven	e-rater-S	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 determine weights for the 12 features.

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 substantively driven automated scores differed markedly. 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 agreeing to participate were assigned to one of two committees in such a way as to create a balance within each committee according to job type and gender.

Each committee consisted of five members, with each committee meeting 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. That process included several iterations in which each member decided on weights individually, the committee engaged in a group discussion around those individual selections, and the members finally revised their weights individually.

The weight-selection process initially separated dimensions and features. That is, each of the five dimensions was weighted on a 0–100 scale and, then, each of the 12 features was weighted on a 0–100 scale. Next, each feature weight was multiplied by its dimension weight. This process intentionally drove the weighting of lower-level features from dimensions that were arguably closer to the intended target construct. In effect, each feature weight was strongly constrained by its dimension weight.[4]

It is also worth noting that, in this process, 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 in the abstract from its perceived importance *given knowledge of how e-rater-H measures it*. This distinction was desirable to capture because e-rater-H's implementation of a dimension may not be what experts mean when they think of that same dimension.

Last, it should be noted that 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 and their acceptability from a substantive perspective. 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 substantively acceptable values. Because of this fact, the second committee chose its features and weights without knowing anything about the optimally predictive values.

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, comparing the automated approaches).

The training sample was used by all three approaches to create the vectors of words that are needed for computing feature scores related to topical analysis. Additionally, the training sample 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.[5]

The training sample consisted of 250 students selected from the 568 WOL students who had participated in the main NAEP 2002 reading assessment. To allow for a representative distribution of scores on both essays, 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 over-sample the tails of the distribution so that there were enough extreme scores to train on.

Four e-rater models were built for each essay: one model for e-rater-E, one for e-rater-H, one from the mean weights set by committee 1 that had knowledge of the e-rater-H empirical weights ("e-rater-S1"), and one based on the mean weights set by committee 2 that did not have knowledge of the empirical weights ("e-rater-S2").

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.

Stage 3

In the third stage, experts evaluated a selected sample of essays for which the first human rating and e-rater-S 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-S1 or e-rater-S2 scores diverged most (by 2–4 points) from the human scores awarded to the same responses.

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-S 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 substantively driven model to other essays.

Two persuasive prompts and 2 informative prompts were selected from the 20 essay prompts used in the 2002 main NAEP writing assesment (but minus the two prompts already taken from this set for the WOL study). The essays from each genre were chosen to be as similar as possible in score distribution to the original study essays and to show a range of variation in terms of the characteristics of the prompts themselves (i.e., whether stimulus material was provided; whether a letter, news article, or traditional essay was called for; whether the task was abstract or 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. To generate topical analysis feature scores for responses to the new prompts, a training sample of 100 responses (out of 300) was used for each new prompt to provide the word vectors. 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-S1, and e-rater-S2 as originally created for the two essays in the WOL data set.[6]

Results

To What Extent Are Judgmentally Determined Weights Reproducible?

The reproducibility of judgmentally determined dimension weights was evaluated across committees and across individuals within a committee.[7] In the current study, there were only two committees and only five members on each one, so the results with respect to reproducibility are at best suggestive of what might occur from other, similarly conducted weighting activities.

Reproducibility across Committees

To evaluate the reproducibility of judgmentally determined dimension weights, the initial mean dimension weights were compared across the two committees, and then the mean of the absolute differences was computed. The initial dimension weights were generated *before* committee members were introduced to the specific features used by e-rater-H to measure the dimensions. For these initial dimension weights, the mean of the absolute differences between the two committees was 4 points for essay 1 (range = 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 substantively based, another measure of reproducibility across committees was 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. e-rater's mean absolute differences ranged from 17–21 points across essays

and committees as compared to the 3–4-point difference between the two committees. Table 4 shows the committee initial dimension weights and e-rater-H dimension weights.

Table 4: 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 the committees).

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. Given these possible changes it is useful to compare the *final* mean dimension weights across committees and also between committees and e-rater-H.

As table 5 (next page) shows, for the final weights the mean absolute differences between the two committees increased somewhat. At the same time, the mean absolute difference between e-rater-H and committee 1 (which was shown the empirical weights) became noticeably smaller. 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 5: 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. 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 committee 1 assigned markedly lower dimension weights than committee 2 on Topical analysis, and notably higher weights than committee 2 on Essay length. 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.

Reproducibility within Committees

Dimension weights across individuals within a committee were far less reproducible than weights across committees, indicating that while the two committees were similar in their aggregated judgments, those aggregations did not always represent a within-group consensus. For committee 1, 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, 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).

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.

Committee members noted that style was missing from both the scoring rubrics and, in large part, from e-rater-H, which doesn't detect such text characteristics as extended metaphor, personal voice, figurative language, rhetorical devices (e.g., purposeful repetition), language sophistication, and unconventional organization. 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. Committee members also thought that audience awareness was missing from both e-rater-H's implementation and from the NAEP rubrics, and that although cohesion was implied by the NAEP rubrics' inclusion of transitions, e-rater-H appeared to take no explicit account of it. With respect to Word complexity, committee members noted that "word choice" (included in the NAEP rubrics) was a more appropriate consideration because more difficult words are not necessarily better ones. Regarding Topical analysis, 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.

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-S. Three categories of analysis were run. These analyses concerned relations with human scores, relations with other indicators, 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 6 (next page) 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 of ratings. As a consequence, the two human scores were averaged to form a more reliable estimate of each examinee's true score (labeled "Human R1 + R2").

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-S2 score was significantly lower than the combined human score for both essay 1 (standardized mean difference, or d, = .20) and essay 2 (d = .21). In addition, e-rater-S1 produced significantly lower scores than the combined human score for essay 2 (d = .11).

Table 6: 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-S1	3.6	1.3	3.6	1.2
e-rater-S2	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-S1	3.3	1.2	3.3	1.2
e-rater-S2	3.2	1.3	3.2	1.2

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

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; however, post-hoc tests showed that more of the machine methods differed from the human method. For essay 1, e-rater-H awarded significantly *higher* scores than the scores given by the first human rating (d = -.06), while e-rater-S2 awarded scores that were significantly lower than that first human rating (d = .16). For essay 2, *all* machine methods produced scores that were significantly lower than the human scores (d range = .06 to .24).

Table 7 shows the intercorrelations among the four automated scoring approaches. Of note is that the e-rater-S1 and e-rater-S2 approaches strongly intercorrelated (r = .86 for essay 1 and .90 for essay 2). Even so, the two methods were different in their relations to the other automated approaches. e-rater-S1's correlation with e-rater-H was higher than e-rater-S2's correlation with e-rater-H for both essays 1 (.92 vs. 81. and .90 vs. .84 for essays 1 and 2, respectively). Also, e-rater-S1's correlation with e-rater-E was significantly higher than e-rater-S2's correlation with e-rater-E (.77 vs. .67 and .81 vs. .74). These differences in functioning between the two e-rater-S approaches can only be due to the feature weights, which constitute the sole distinction between them.

Table 7: Intercorrelations among the Automated Essay Scoring Approaches for the Total Cross-Validation Sample (N = 1,055)

	Essay 1			Essay 2		
	e-rater-E	**e-rater-H**	**e-rater-S1**	**e-rater-E**	**e-rater-H**	**e-rater-S1**
e-rater-H	.75	—		.77	—	
e-rater-S1	.77	.92	—	.81	.90	—
e-rater-S2	.67	.81	.86	.74	.84	.90

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

Percentages of exact agreement among the four automated approaches revealed a similar pattern of association to that depicted by the correlations. (See Bennett & Ben-Simon, 2006, for complete results).

Table 8 gives the correlations of each e-rater approach with the first human rating and with the mean of the two human ratings. As the table shows, the correlations between e-rater-S1 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 S1 scores were correlated a few points higher with the human scores than the e-rater-E scores correlated with the human scores. In contrast, the e-rater-S2 scores correlated consistently less well with humans than did the e-rater-H scores for both essays or than did the e-rater-E scores for essay 1. The differences in functioning between the two versions of e-rater-S derive from their feature weights. For S1, these weights were closer to the optimal, empirically derived weights used by e-rater-H.

Table 8: 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-S1	e-rater-S2
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.

Percentages of exact agreement between each automated approach and the human ratings were also calculated. e-rater-S1's exact agreement with the human ratings 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. e-rater-S2's agreement ran between 6 and 14 points lower than e-rater-H's values and 8 to 12 points lower than e-rater-E's values.

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 9 gives the results.

Table 9: Correlations between Essay 1 and Essay 2 Scores for the Total Cross-Validation Sample (N = 1,055)

Scoring Method	Correlation between Essays	T-Test values
Human R1	.61	—
e-rater-E	.54*	$t_{1002} = 3.24, p < .01$
e-rater-H	.64	n.s
e-rater-S1	.63	n.s
e-rater-S2	.55*	$t_{1002} = 2.77, p < .01$

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

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 and e-rater-S2, both of which had cross-essay correlations lower than the human value.

Relations with Other Indicators

The analyses in this section explore the extent to which the different automated methods can be distinguished in their relationships to two other indicators. Among the measures in the WOL data set were main NAEP writing performance information and the number of words comprising each essay.

Writing performance information was available for that subset of the cross-validation sample taking the main NAEP writing assessment. For those students, this information takes the form of "plausible values." These plausible values represent five random draws from an estimated ability distribution based upon student responses to the test, 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 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.

The second measure to be considered is the number of words comprising each essay. Even though essay length is explicitly represented in e-rater-H's and e-rater-S'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 also implicitly represented through other features.

Shown in Table 10 are the correlations between each scoring approach and the two other 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 was significantly lower than the correlation between the first human rating and main NAEP performance. Second, *all* of the automated methods correlated significantly more strongly with essay length than did the first human rating.

With respect to essay length, e-rater-S1 was more related to this feature than was e-rater-H. This higher relationship occurred even though e-rater-S1'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-S1 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 or e-rater-S2.

Table 10: Correlations between the Scoring Approaches and Two Other Indicators for the Cross-Validation Sample

Indicator	N	Human R1	e-rater-E	e-rater-H	e-rater-S1	e-rater-S2
Essay 1						
Main NAEP writing[a]	687	.52	.46*	.49	.49	.44*
Essay length	1,005	.57	.74*	.81*	.87*	.73*
Essay 2						
Main NAEP writing[a]	687	.56	.47*	.53	.53	.52
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 correlation reported is the average correlation (using the Z-score transformation) between the rating and each of five plausible values.

Resolution of Large Human-Machine Score Discrepancies

For these analyses, a sample of 60 responses to each of the two essays was examined for which the human and e-rater-S scores differed markedly. To help identify whether the expert committees found the e-rater-S scores more or less acceptable relative to human scores, each committee member was given the discrepant responses resulting from the application of e-rater-S with that committee's weights. Committee members were also given the first human rating and the e-rater-S scores. For each discrepant response, committee members were asked to choose blindly the more appropriate score (human or e-rater-S) 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 11 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 generally 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 11: Correlations between Mean Resolved Scores of Committee Members and Automated Essay Scoring Approaches

	Human R1	e-rater-E	e-rater-H	e-rater-S
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: A separate sample of 60 responses was selected for each committee and essay. Committee 1 reviewed discrepant responses for e-rater-S1 and committee 2 reviewed discrepant responses for e-rater-S2.

There were also differences among the automated approaches in their relations with the resolved scores. For both essays, the e-rater-H scores and the e-rater-E scores correlated noticeably higher with the resolved scores than the e-rater-S2 scores correlated with the resolved scores. Last, for essay 1 the e-rater-H scores correlated higher with the resolved scores than did the e-rater-S1 scores. (These comparisons of e-rater-S to the other automated approaches need to be viewed cautiously as the included responses were chosen because *e-rater-S* – and not the other approaches – scored them discrepantly.)

Table 12 (next page) shows means and standard deviations 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 are also indicated. The statistical test was a repeated-measures ANOVA conducted separately for each essay and version of e-rater-S, with scoring method as the independent variable and essay score as the dependent variable.

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-S 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. 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 12: Means and Standard Deviations 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-S	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	0.9	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-S1 and committee 2 reviewed discrepant responses for e-rater-S2. Human R1 = first human rating.

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

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 (69%-76%). The remaining three categories were of secondary or, sometimes, negligible importance.

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-S 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 reasons suggesting instances in which the student's response was simply a restatement of the prompt that was scored higher by e-rater-S 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 Substantively Driven Scoring Model Developed for One NAEP Prompt Generalize to Other NAEP Prompts of the Same Genre?

To address this question, the e-rater-S 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-S scoring model created for grading the persuasive prompt (Essay 2) was employed for scoring two new prompts from that genre. 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 correlations of the automated approaches with the first human rating (table 13, next page). Surprisingly, the correlations did not appear to have attenuated appreciably from those observed for the original essays. Also, across all essays and samples, e-rater-S1 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-S2, however, the correlation with the first human rating was lower for three of the four new essays than was the correlation of e-rater-H with the human ratings. e-rater-S2 was less related to human scores than was e-rater-S1 only for the two informative essays.

Table 13: Correlations of the Automated Scoring Approaches with the First Human Rating for the Original Essays in the Cross-Validation Sample (1005) and for New Informative and New Persuasive Essays in the Generalization Samples

	e-rater-E	e-rater-H	e-rater-S1	e-rater-S2
Informative				
Original Essay 1 (N = 1005)	.66	.67	.66	.59
New Essay 1 (N = 200)	.66	.75	.73	.58
New Essay 2 (N = 198)	.67	.67	.65	.50
Persuasive				
Original Essay 2 (N = 1005)	.69	.72	.73	.68
New Essay 1 (N = 199)	.60	.62	.59	.59
New Essay 2 (N = 200)	.67	.71	.68	.63

Discussion

The objective of this study was to lay the groundwork for a more substantively 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 dimen-

sions 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, to other indicators, and in the resolution of large machine-human discrepancies. The third study question focused on how well the substantively driven scoring model developed for one NAEP prompt generalized to other NAEP prompts of the same genre. To address these questions, two versions of the substantively driven 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 substantively based version derived from the committee that was aware of the hybrid's weights was dubbed e-rater-S1. The version derived by the committee independently of knowing the hybrid's weights was called e-rater-S2.

Table 14 summarizes the results with respect to the two study questions. The table includes only those analyses that showed consistent differences in functioning for e-rater-S scores (where "consistent" was defined to mean a similar result across both essays for the second study question and across at least three of the four essays for the third, or generalization, study question). As can be seen, there are few consequential differences between e-rater-S1 and the other automated approaches. In contrast, e-rater-S2 showed many consistent differences in functioning, some of which were quite substantial.

Table 14: Consistent Differences between e-rater-S and Other Automated Approaches

Analysis	e-rater-S1	e-rater-S2
Relations with Human Scores		
Mean differences		• S2 < human by .16 – .24 SD units
Correlations with human scores		• S2 < hybrid by .04 – .08 points
Percentage of exact agreement with human scores	• S1 < hybrid by 3 – 7 points • S1 < empirical by 2 – 6 points	• S2 < hybrid by 6 – 14 points • S2 < empirical by 8 – 12 points
Inter-prompt correlations		• S2 < human by .06 points
Relations with Other Indicators		
Correlation with Essay length	• S1 > than other automated approaches • S1 > than human	• S2 < than other automated approaches • S2 > than human
Large Machine-Human Score Discrepancies		
Correlations with resolved scores		• S2 < hybrid by .35 and .19 points • S2 < empirical by .30 and .27 points
Generalization Analysis		
Correlations with human scores		• For three of four essays, S2 < hybrid by .08-.18 points

Note: Empirical = e-rater-E. Hybrid = e-rater-H. Only analyses showing consistent differences in functioning for e-rater-S scores are included (i.e., a similar result across both essays for relations with human scores, relations with other indicators, and large machine-human discrepancies, and across at least three of the four essays for the generalization analysis).

Overall, then, the results seem to suggest that the two versions of the substantively driven approach operated differently from one another. e-rater-S1, 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-S1 and e-rater-H scores were in the low .90s for both of the two original essays.) The e-rater-S1 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 substantive 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 substantive 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 substantive 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 necessarily 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 a more accurate representation of how effectively the nation's students write, NAEP is scheduled to include measures of writing on computer in its 2011 assessment (Olson, 2007). 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.[8]

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 (a) a comprehensive expert analysis of the extent to which the features as implemented adequately cover the dimensions derived from the theory and (b) 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. It is not clear whether other automated scoring programs – or even more recent releases of e-rater – 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-S and the other two approaches. The two versions of e-rater-S, 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.

Endnotes

1. The description of e-rater v2 in Attali and Burstein (2005) differs from the description given in Attali and Burstein (2006). In the latter version, only 10 of the original 12 features are included. In the current study, we use the 2005 description because that was the version used operationally by ETS at the time of this study.
2. The term "informative" is used by NAEP to denote a genre of writing that "communicates information to the reader to share knowledge or to convey messages, instructions, and ideas" (NCES, 2004).
3. e-rater v1.3 and v2.1 are those modifications of e-rater v1 and v2, respectively, in use at the time this study was conducted.
4. 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.
5. This scaling was done somewhat differently for each approach. 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. These differences would appear to have had only negligible impact on results. See Bennett & Ben-Simon (2006) for details and analysis.
6. 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.
7. The reproducibility of feature weights was evaluated also and is presented in Bennett and Ben-Simon (2006). Because the dimension weights strongly constrain the feature weights, only the former analysis is presented here.
8. 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

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

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater version 2.0. *Journal of Technology, Learning and Assessment, 4*(3). Retrieved May 20, 2007 from http://www.bc.edu/research/intasc/jtla/journal/pdf/v4n3_jtla.pdf

Bennett, R.E. (2006). Moving the field forward: Some thoughts on validity and automated scoring (pp. 403–412). In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing.* Hillsdale, NJ: Erlbaum.

Bennett, R.E., & Ben-Simon, A. (2006). *Toward theoretically meaningful automated essay scoring* (Report 329). Jerusalem, Israel: National Institute for Testing and Evalution. Retrieved May 20, 2007 from http://clickit.ort.org.il/files/upl/224004906/934404432.pdf

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.

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 IntelliMetricTM 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.

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, & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2). Available: http://escholarship.bc.edu/jtla/vol5/2/

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

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

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 Center for Education Statistics (NCES). (2004). *What does the NAEP writing assessment measure?* Retrieved May 20, 2007 from http://nces.ed.gov/nationsreportcard/writing/whatmeasure.asp

Olson, L. (2007). NAEP writing exams going digital in 2011. *Education Week, 26*(27), 23.

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., 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.

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.

Acknowledgments

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

We are grateful to the following colleagues for their assistance: Yigal Attali, Jill Burstein, Chi Lu, Fred Yan, Bruce Kaplan, Mark Shuvman, Henry Braun, Shelby Haberman, Neil Dorans, Charlie Lewis, Dan Eignor, Don Powers, Marcia Ashhurst-Whiting, Anthony Bucco, Gail Hawisher, Geof Hewitt, Brian Huot, Tanji Reed Marshall, Brian Medley, Patricia McGonegal, and Lee Odell.

Author Biographies

Anat Ben-Simon is a Senior Research Scientist, and director of the department of special projects at The National Institute for Testing and Evaluation (NITE) in Jerusalem. She obtained her doctorate in cognitive psychology and psychometrics from the Hebrew University of Jerusalem. At NITE, Dr. Ben-Simon has headed the Department of Computer-Based Testing, and directed the Israeli National Assessment of Educational Progress. Currently Dr. Ben-Simon is directing a number of projects, including the development of MATAL, a computerized test battery for the diagnosis of learning disabilities among post secondary students, and the Hebrew Language Project (HLP), which involves the development of tools for automated essay scoring and text analysis. She also holds a teaching position in the Psychology Department of the Hebrew University of Jerusalem. Anat Ben-Simon can be contacted at anat@nite.org.il.

Randy Elliot Bennett is Distinguished Scientist in the Research & Development Division at Educational Testing Service in Princeton, New Jersey. A graduate of Teachers College, Columbia University, Dr. Bennett began his employment at ETS in 1979. Since the 1980s, he has conducted research on the applications of technology to testing, on new forms of assessment, and on the assessment of students with disabilities. Dr. Bennett's work on the use of new technology to improve assessment has included research on presenting and scoring open-ended test items on the computer, on multimedia and simulation in testing, and on generating test items automatically. For this work, he was given the ETS Senior Scientist Award in 1996 and the ETS Career Achievement Award in 2005. He is the author of many publications including "Technology and Testing" (with Fritz Drasgow and Ric Luecht) in Educational Measurement (4th Edition) and "What Does it Mean to Be a Nonprofit Educational Measurement Organization in the 21st Century" (http://www.ets.org/Media/Research/pdf/Nonprofit.pdf). Randy Elliot Bennett can be contacted at rbennett@ets.org.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

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