ED 458 296	TM 033 472
AUTHOR	Kelly, P. Adam
TITLE	Computerized Scoring of Essays for Analytical Writing Assessments: Evaluating Score Validity.
SPONS AGENCY	Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
PUB DATE	2001-04-00
NOTE	43p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Computer Assisted Testing; Elementary Secondary Education; *Essays; Higher Education; Scores; *Scoring; *Test Scoring Machines; *Validity; *Writing Evaluation
IDENTIFIERS	*Analytical Writing

ABSTRACT

The purpose of this research was to establish, within the constraints of the methods presented, whether the computer is capable of scoring essays in much the same way that human experts rate essays. The investigation attempted to establish what was actually going on within the computer and within the mind of the rater and to describe the degree to which these processes equated. Revealing this parallelism depended on the careful assessment of the "intrinsic" aspects of validity as proposed by S. Messick (1995) of computerized essay scores. The focus was "e-rater" (TM), a computer-based essay scoring system developed by the Educational Testing Service. The study used 1,794 existing Graduate Record Examination Writing Assessment essays written and scored during recent test administration. Factor analysis and the advice of expert raters were used to guide the deconstruction of essay scoring models into subscore models corresponding to writing characteristics within the essay. The writing characteristics identified in this process were used as the basis for developing characteristic-specific scoring rubrics to be used by expert raters. Fresh essay samples were scored by expert rates, both holistically and characteristic-by-characteristic. The same essay samples were assigned both holistic scores and character-wise subscores by the computer scoring models. The degree of convergent validity of scores was evidenced by the proportion of agreement and strength of pairwise correlation among scores and subscores. The statistics derived in this study suggest that simpler e-rater models might do just as well at agreeing with the scores of expert rates, although the proportion of total variance in the expert rater scores explained by the e-rater scores might decrease from an already modest level. (Contains 7 tables and 30 references.) (Author/SLD)

Computerized Scoring of Essays for Analytical Writing Assessments:

Evaluating Score Validity

P. Adam Kelly

Department of Educational Research
College of Education
Florida State University

Paper presented at the Annual Meeting of the National Council on Measurement in Education

April 2001

Acknowledgments

This research is supported by the Graduate Record Examination Board and its Research Committee. The author would like to thank the following people: Tom Rochon, Executive Director of the GRE Program, and the other members of the GRE Board, for their generous support of this study; Dr. Don Powers, Educational Testing Service, for playing a key role in obtaining the support of the GRE Board, providing a review of the first draft, and serving as mentor throughout the project; Dr. Jill Burstein, ETS Technologies, Inc., for providing the *e-rater* data necessary to carry out this research and reviewing the first draft; and Mary Fowles, ETS, for recruiting the expert raters, scheduling the face-to-face components of the project, editing and re-editing the study materials, and reviewing the first draft.

Also, thanks to Cyndi Welsh for helpful comments in reviewing the study materials, Cynthia McAllister for designing the scoring sheets and managing the correspondence, Slava Andreyev and Chi Lu for programming and running the many *e-rater* models, Irv Katz for observing and providing advice on the "talk-aloud" procedures, Ann Willard for helpful comments on the first draft, Dr. Eiji Muraki for providing invaluable assistance in laying out the factor analysis results, and Profs. Donna Tobin, Maria Cimadevilla, Marian Kerr, and Joyce Munro, for patiently and professionally carrying out the rating tasks, at times rigorous, that were requested of them.

Thanks also to Dr. Albert Oosterhof, Florida State University, for providing helpful advice and moral support as Major Professor, and to Dr. Akihito Kamata, Dr. Richard Tate, and Dr. Myke Gluck, all of Florida State University, for their help as members of the Dissertation Committee.

Abstract

The purpose of this research is to establish, within the constraints of the methods presented, whether the computer is capable of *scoring* essays in much the same way that human experts *rate* essays. That is, this investigation attempts to establish what is actually going on within the computer *and* within the mind of the rater, and describe to what degree these processes equate. The essence of how this parallelism is to be revealed lies in the careful assessment of the "intrinsic" aspects of validity, as proposed by Messick (1995), of computerized essay scores.

Factor analysis and the advice of expert raters are used to guide the deconstruction of essay scoring models into subscore models corresponding to writing characteristics within the essay. The writing characteristics identified in this process are also used as the basis for developing characteristic-specific scoring rubrics to be used by expert raters. Fresh essay samples are then scored by expert raters, both holistically and characteristic-by-characteristic.

These same essay samples are assigned both holistic scores and characteristic-wise subscores by the computer scoring models. The degree of convergent validity of scores is evidenced by the proportion of agreement and strength of pairwise correlation among the scores and subscores.

Table of Contents

Introduction

The ever-increasing quantity of large-scale testing, coupled with the growing movement towards more direct forms of student assessment, creates a huge burden on test scoring resources. Specifically, the rise in constructed-response assessment leads to a greatly increased need to evaluate many more responses efficiently. Figure 1 shows the growth in volume of several major essay tests over the last decade. Many in educational testing believe that the time has come for computerized essay scoring tools, envisioned as reliability checks – but *not* replacements – for expert raters.

Recent studies have shown a high degree of correlation between expert rater scores and computer-generated scores from several computerized scoring tools. While such evidence of reliability is essential to the success of these tools, it is equally essential that the tools generate scores that are credible in their own right. That is, in scoring an essay, a computerized scoring tool should ideally engage in processes that can be shown to parallel, if not exactly duplicate, what an expert rater would do. The degree to which these tools are accepted will likely depend on such parallels being demonstrated convincingly.

This study presents and evaluates validity evidence of the interpretation and use of scores produced by computerized essay scoring tools, following the test score validity reasoning of Messick (1987, 1995). Applied in the current context, the premise is that the interpretation and relevance/utility of computer-generated essay scores should be no different from the interpretation and relevance/utility of expert rater-generated scores. Specifically, this study

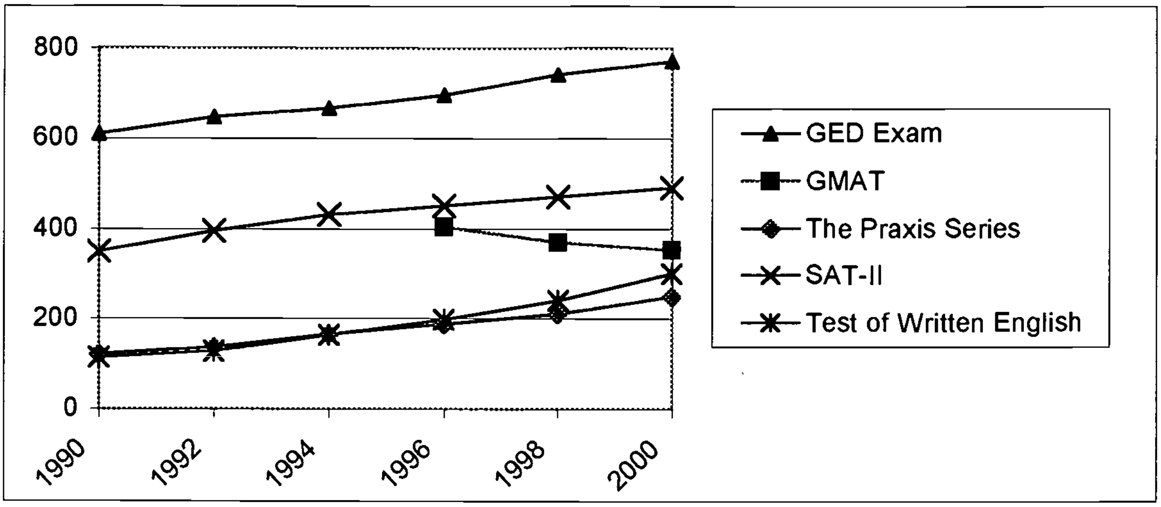

Figure 1. Growth in Selected Essay Tests (in thousands of essays), 1990-2000. Statistics provided by the American Council on Education, The College Board, Educational Testing Service, and the Graduate Management Admissions Council.

assesses whether a claim can be reasonably made that a computerized essay scoring tool takes into account many of the same things that an expert rater does in determining an essay score. At present, the apparent high correlation of computer-generated scores with expert rater-generated scores remains unexplained in that no empirical linkage of the electronic processing of a computer to the cognitive processing of a human being has been established for the essay scoring process.

The focus of this study is *e-rater*™, a computer-based essay scoring system developed by Educational Testing Service® (ETS®). In 1999, *e-rater* was implemented as the "second rater" on essays from administrations of the computer-based Graduate Management Admissions Test® (GMAT®). Each GMAT essay is scored by an expert rater and by *e-rater*. If the two scores differ by more than one point, the essay is sent to a second expert rater for adjudication. In addition to being used for operational GMAT scoring, *e-rater* is also used in the evaluation of essays submitted through Criterion^{SM} Online Writing Evaluation Service, offered by ETS Technologies, Inc., a for-profit subsidiary of ETS. *E-rater* is presently one of four computerized essay scoring services available commercially; the others are Project Essay Grade™ (PEG™), offered by Tru-Judge, Inc., Intelligent Essay Assessor™ by Knowledge Assessment Technologies, and IntelliMetric™ by Vantage Learning.

Background

The reasons for the move towards greater use of essays are evident. Essays address cognitive constructs unique to written communication, such sustaining a well-focused, relevant discussion, that may not be plausibly measured with less direct item types (Powers, Burstein, Chodorow, Fowles, & Kukich, 2000; Braun, 1988). Additionally, the notion that essays *appear* more reflective of communication skills desirable in an academic setting has spurred the growth of essay use. The issue is not even so much whether it is *right* to use essays in a test, but simply that essays *are* used, and their use is defended locally with a variety of justifications.

There are many reasons given against the use of essays, as well, including difficulty in defining the constructs measured in an essay test and the score validity problems that arise from interrater unreliability. A universal criticism, however, is the high cost of scoring essay tests. As the number and complexity of essay tests increase, the pressure increases on administrators of these tests to get, train, and retain raters; on funding sources to cover the costs of attracting the highest-quality raters; and on professors' and teachers' time outside the classroom available for rating which, after all, is a second job for most raters. Moreover, the situation may worsen as the culture of direct assessment – and the technology to promote it – both continue to grow (Messick, 1999).

A solution may be to have computers score essays. This offers several advantages:

- Cost Savings: If used alone, a computerized scoring tool eliminates nearly all of the cost of scoring an essay. As currently used on the GMAT Analytical Writing Assessment, as a "reliability check" following an initial rating by an expert rater, it cuts the cost of scoring significantly.

- Reliability: A (high-quality) computer never loses its virtually incorruptible consistency in scoring.

- Accessibility: As access to technology grows, computerized scoring of essays becomes increasingly available.

A big question is, are the scores generated by computerized scoring tools valid? This question has been addressed in several ways to date:

- ♦ At ETS, multiple studies have shown that *e-rater*-generated essay scores agree frequently – between 88 and 97 percent of the time – with scores produced on the same essays by expert raters, under a variety of conditions (e.g., Burstein, Kukich, Wolff, Lu, & Chodorow, 1998a, 1998b; Kaplan, Wolff, Burstein, Lu, Rock, & Kaplan, 1998; Burstein & Chodorow, 1999). The proportion of agreement has been shown to be particularly high when "agreement" is defined as either two identical scores or a difference of one point between scores, a commonly accepted definition. Other ETS research has suggested that *e-rater*-generated scores correlate nearly as well with non-test indicators of writing skill as expert rater scores do (Powers, et al., 2000), and that at least one free-response test– the writing assessment portion of the National Assessment of Educational Progress (NAEP) – provides evidence of some identifiable, construct-centered structure underlying *e-rater* scores (i.e., a tying of the scores to constructs like discourse, syntactic variety, and on-topic content) (Muraki, Lee, & Kim, 2000).

- Similar research has been undertaken with PEG: Correlation studies (e.g., Page, 1994; Page & Petersen, 1995) show a consistently higher correlation between a PEG score and the score of any single rater, or any mean of two or three raters, than any single rater or multiple-rater mean has with another; and trait-scoring studies (Page, Poggio, & Keith, 1997; Shermis, Koch, Page, Keith, & Harrington, 1999; Keith, 1999) provide evidence of relatedness between underlying traits of writing and the overall PEG score assigned to an essay.

Messick (1987) identifies and examines the interaction of evidential and consequential bases of test score validity with proposed test score interpretations and uses. He most succinctly represents his view of test score validity with his 2x2 matrix, shown in Figure 2.

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U +VI + Social Consequences

Figure 2. Messick's (1987) Construct Validity Matrix.

Messick (1995) elaborates on a procedure for investigating the validity of test scores for particular interpretations and uses. He provides a "road map" for his 2x2 matrix, in the form of six "distinct aspects [of construct validity] to underscore issues and nuances that might otherwise be downplayed or overlooked ... [to] function as general validity criteria or standards for all educational or psychological measurement" (Messick, 1995, p. 744). These six aspects are:

> Content: Evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989b);
>
> Substantive: Theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks;
>
> Structural: Appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957; Messick 1989b);
>
> Generalizability: Examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test criterion relationships (Hunter, Schmidt, & Jackson, 1982);
>
> External: Includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965);
>
> Consequential: Appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989b).
> (Messick, 1995, p. 745)

The investigation presented here covers the first three of these, the "intrinsic" aspects of test score validation, while a companion investigation, currently in progress, covers the latter three, "extrinsic" aspects. The results of both investigations are to be reported in the researcher's dissertation, scheduled for completion in mid-Summer, 2001.

Procedures

All data, essay scoring rubrics, and *e-rater* models were provided by ETS and ETS Technologies, Inc. The study utilized existing Graduate Record Examination® (GRE®) Writing Assessment essays written and scored during recent test administrations. The procedures were performed in three phases, paralleling the first three aspects of Messick's (1995) validation technique. Briefly, *content relevance and representativeness* were examined through the identification and assessment of the factor structure of each GRE Writing Assessment essay type. The results of this assessment were used in the construction of characteristic-specific holistic scoring rubrics. *Reflectivity of the task and domain structure* was assessed through comprehensive reviews, conducted by ETS experts in essay test development, of the characteristic-specific rubrics. These experts attended particularly to assessing whether the rubrics target the desired constructs, and not merely the performance of rote counting tasks. Lastly, the *engagement of substantive theories and process models* was evaluated using the results of "talk-aloud" protocols produced from recordings of expert raters rating of a sample of essays. (A detailed description of the procedures followed accompanies the Phase III results.)

The Context: The GRE Writing Assessment: As specified by the GRE Writing Test Advisory Committee (http://www.gre.org/stuwrit.html#description), the GRE Writing Assessment is designed to measure the ability to:

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

The GRE Writing Assessment consists of two tasks: a 45-minute "issue" task that requires the examinee to present a perspective on an issue, and a 30-minute "argument" task that requires analysis of an argument presented. The two tasks are intended to complement each other, in that the first task requires an examinee to construct an argument and the second to critique an argument already made (GRE Website, http://www.gre.org/twotasks.html).

The GRE Writing Assessment essays are scored on a six-point scale, using holistic scoring rubrics. Expert raters, who are college and university faculty, usually in English, Communications, or a discipline within the humanities (e.g., Rhetoric and Composition, English Literature, Classics, Linguistics), are trained to evaluate GRE essays over a two- to three-day period and then also participate in various rating norming procedures throughout each scoring session. For example, one or more times per day during a scoring session, raters take part in a "rangefinding" procedure, in which all raters examine and discuss exemplar essays for each of the six score levels. The intent of these procedures is to ensure that all raters are looking at similar qualities and keeping in mind certain issues as they rate essays. Currently, most rating on the GRE Writing Assessment is done on-line from the rater's home location, with instant, on-line support from a GRE scoring leader.

The Scoring Tool: *E-rater*: Briefly, *e-rater* uses natural language processing techniques to model the performance of expert raters. For each essay prompt, a sample of essays, previously scored by expert raters, is selected such that an adequate number of essays representing each score category is included. These essays are then used to "train" *e-rater* to score new, unscored essays (Burstein, Kukich, et al., 1998a). In its attempt to model expert raters, *e-rater* first uses several subroutines to extract a variety of features from an essay. In the "training" step, these features are used in combination to "postdict" the score previously assigned the essay by expert raters. The system is adjusted when necessary to provide the maximum agreement between the *e-rater* "postdicted" scores and the actual expert rater scores in the model-building sample. The system is then ready to score new essays.

Powers, et al. (2000) provide a short overview of *e-rater*, the main portion of which is summarized here. Currently, the *e-rater* system is built on a stepwise ordinary least squares (OLS) regression model. *E-rater* focuses on three general classes of essay features: *discourse*, indicated by various rhetorical features that are expected to occur throughout an essay; *syntactic*, indicated by the structure of sentences; and *content*, indicated by prompt-specific vocabulary expected to be present in the essay. A total of 59 features are "extractable," but in practice usually only the most predictive features, as measured by their regression weights, are retained and used for further scoring. The features used must be both predictive of expert rater scores *and* analogous in some recognizable way to the characteristics that expert raters are trained to consider. The OLS regression weights for these features are applied to each new essay to estimate a score. The estimated score is then rounded to the nearest integer, from 0 to 6, in order to make its scale conform to that used by expert raters.

The Sample: The GRE Program provided a total sample of 1,794 GRE Writing Assessment essays, consisting of 620 essays written on the "issue" prompt type, and 1,174 essays written on the "argument" prompt type. ETS Technologies' staff divided the "issue" essay sample into a model-building sample of 226 essays and a cross-validation sample of 394 essays; the "argument" sample was divided into 251 model-building essays and 923 cross-validation essays. All of these essays are considered "operational," that is, written by actual graduate school candidates under ETS' standard testing conditions, and previously scored by two expert raters.

(In operational scoring, the GRE Writing Assessment is always scored by two expert raters; the implementation of a computerized scoring tool is not under consideration.)

Results

Phase I: Content Relevance and Representativeness: The first phase of the study aims to evaluate how the structure of the *e-rater* model addresses the content and construct domains of the GRE Writing Assessment. The evaluation consists of measuring and mapping the 59 *e-rater* features to a set of factors, or underlying writing characteristics. To accomplish this, exploratory factor analyses of the 59 features is performed on the model-building sample data for each of the two essay types. Powers, et al. (2000) reason that there is some advantage to employing a "generic" *e-rater* scoring model that spans multiple prompts within an essay type. First, as a matter of practicality, a generic model may be "trained" using any essay of its essay type available, regardless of the prompt on which the essay was written. This makes for more convenient "training" of models. Second, as a matter of score validity, generic-model scores may provide some supporting evidence for *e-rater* score generalizability for an essay type, in the sense that all essays of a particular type are scored by the same model and, therefore, should reflect the essay type rather than the individual prompt. The present study examines the evidence of this score generalizability argument by performing principal component analyses on data for both a "generic" model and three prompt-specific models separately for each essay type and comparing the results.

The data were analyzed using the principal component and principal factor methods, following a procedure established by Muraki, et al. (2000) for analyzing NAEP essay data processed by *e-rater*. The principal component analysis was used first to identify the features, out of the set of 59 included in *e-rater*, that account for the largest variations in *e-rater* scores. ETS Technologies classifies the features into three categories: discourse (rhetorical), syntactic, and content. Forty-five features are classified as discourse, eleven as syntactic, and three as content. After the most explanatory features were selected, principal factor analysis was used to estimate the structure of these features.

The first 20 principal components explained as a whole approximately 81 percent of the total variance in the model-building sample data sets for both the "issue" and "argument" prompt types. The first component accounts for over 17 percent of the total variance in *e-rater* scores for the "issue" prompt type, and nearly 16 percent for the "argument" prompt type. For both prompt types, each of the next four components explains at least four percent of score variance. Beyond the fifth component, the contribution of additional components to explained variance is steady but small. Overall, about half of the total score variance can be explained by the first six components for each prompt type. (These results closely parallel what was found by Muraki, et al. (2000).)

Based on the results of the principal component analyses, a total of 25 features that had loadings larger than 0.5 for the "issue" prompt type, and 21 features for the "argument" prompt type, were selected for further modeling. These features were selected because they contributed to the first four principal components. There were very few loadings in excess of 0.5 for any feature beyond the fourth component, and none at all for the fifth through the seventh components. The same principal components analysis was performed on subsets of the model-building sample data corresponding to each of the three individual prompts for each of the two essay types. While the subset sample sizes were too small to allow for the drawing of firm conclusions, it appeared that the general pattern described above for the generic models was present in each of the prompt-specific models: four or five components explaining just under half of total score variance, and 20-25 features loading highly on these first four or five components.

In the principal factor analysis, the features selected from the principal components analyses were fitted to a variety of factor models, and the solutions were rotated in an attempt to improve interpretability. For both the "issue" and the "argument" prompt types, a six-factor model, rotated by the promax method, provided the most interpretable solution. As discussed shortly, there appeared to be a substantial degree of dynamic interaction of two factors, manifested in the form of "echoing," of loadings (i.e., repetition of loadings greater than 0.5 for a given feature) across factors. A way of accommodating this in the factor interpretations is to boost the "kappa" inter-factor correlation coefficient in the promax method. In this analysis, the kappa setting for both the "issue" and "argument" factor models was raised from the default setting of 4 to a setting of 25.

As an added precaution to ensure that valuable information had not been discarded by reducing the number of features following the principal components analysis, similar six-factor models, rotated by the same promax method, were fitted to the original, full set of 59 features. For both the "issue" and "argument" models, a similar pattern feature loadings emerged, although the relative positioning of the factors shifted somewhat. However, there was no evidence that valuable information on the underlying structure of the scoring was lost by excluding the non-contributory features.

The factor loadings for the "issue" and "argument" models are presented in Tables 1 and 2, respectively, and the correlations among the oblique axes for the models are presented in Tables 3 and 4, respectively. There appears to be some degree of "echoing" between factors 1 and 4 in the "issue" model and between factors 1 and 5 in the "argument" model: the discourse-related features load most highly on factor 1, while the syntactic features load most highly on factor 4, but both types of features load more than 0.5 on both factors. For this reason, factor 1 in both models is identified as being "dominated" by the discourse-related features, while factor 4 in the "issue" model and factor 5 in the "argument" model are identified as being "dominated" by the syntactic features. The "echoing" of loadings of some syntactic features with some rhetorical features suggests that there might be dynamic relationships between these groups of features in an essay. The syntactic features are included as a measure of syntactic variety, but it may be that certain syntactic structures tend to co-occur more frequently in the expression of certain rhetorical relations or the making of certain types of arguments.

The fit of the six-factor models to the data was assessed by two different methods. First, using the same exploratory procedure as was used on the model-building samples but specifying a six-factor solution a priori, a factor model was fitted to the cross-validation sample for each essay type. Then a comparison of the model-to-data fits of these models to fits of the respective model-building sample models was examined. Second, a confirmatory factor model based on the best-fitting exploratory factor model was fitted to each cross-validation sample, and the fits of these confirmatory models was assessed.

Table 1

Summary of Results for the Principal Factor Analysis: Promax Factor Loadings for the Estimation Data (Kappa = 25)

Issue Generic Model (25 features)

	Factor					
Feature	1	2	3	4	5	6
DIVISOR_1	0.068	*-0.932*	-0.157	-0.029	0.128	0.179
DIVISOR_2	0.369	*0.816*	0.337	0.411	0.456	-0.052
DIVISOR_3	0.073	0.155	*0.938*	0.029	-0.107	0.081
DIVISOR_4	0.173	0.174	*0.956*	0.071	-0.118	0.016
DIVISOR_5	-0.135	*-0.810*	-0.096	-0.121	0.042	0.188
DISCRSE_1	0.296	-0.342	-0.253	0.247	**0.784**	0.138
DISCRSE_2	0.209	0.486	-0.101	0.254	**0.766**	-0.038
DISCRSE_3	**0.928**	0.188	0.319	*0.809*	*0.515*	-0.094
DISCRSE_4	0.266	0.209	**0.921**	0.154	-0.136	-0.105
DISCRSE_5	**0.598**	*0.760*	0.357	*0.578*	0.378	-0.152
DISCRSE_6	**0.815**	0.251	0.117	0.448	0.236	-0.023
DISCRSE_7	**0.612**	0.168	0.234	0.280	0.409	0.032
DISCRSE_8	**0.708**	0.037	0.261	0.498	0.225	0.080
DISCRSE_9	**0.603**	-0.017	0.262	0.400	0.319	0.048
DISCRSE_10	0.390	0.152	0.099	*0.764*	0.346	-0.214
DISCRSE_11	0.357	0.181	0.013	*0.735*	0.297	-0.012
DISCRSE_12	**0.523**	0.187	0.348	*0.556*	0.338	-0.320
SYNTAX_1	*0.806*	0.195	0.062	0.442	0.260	0.057
SYNTAX_2	*0.575*	0.093	0.190	**0.621**	*0.755*	-0.001
SYNTAX_3	*0.778*	-0.039	0.205	0.324	0.455	0.097
SYNTAX_4	*0.554*	0.106	0.161	**0.890**	0.337	-0.139
SYNTAX_5	*0.771*	0.040	0.298	**0.688**	*0.634*	0.048
CONTENT_1	0.195	-0.349	-0.093	0.032	0.486	**0.757**
CONTENT_2	0.022	-0.178	0.061	-0.126	-0.018	**0.742**
CONTENT_3	0.060	-0.012	0.018	-0.082	0.057	**0.859**
Characteristic	ISSUE-1	--	ISSUE-2	ISSUE-3	ISSUE-4	ISSUE-5

□ = Factor-feature bundle (major feature loadings in **bold**)

▒ = Echo bundle (major feature loadings in *italics*)

░ = Features used as divisors only, not used in essay scoring (therefore disregarded in factor naming)

▨ = Divisor feature-dominated factor (disregarded in balance of study)

Note. Tables 1-4 adapted from Muraki, Lee, & Kim (2000). Reproduced with permission.

Table 2

Summary of Results for the Principal Factor Analysis:
Promax Factor Loadings for the Estimation Data
(Kappa = 25)

Argument Generic Model
(21 features)

Feature	Factor 1	 2	 3	 4	 5	 6
DIVISOR_1	0.200	*-0.843*	-0.150	0.154	0.218	0.278
DIVISOR_2	0.296	*0.913*	0.124	0.054	0.264	0.093
DIVISOR_3	-0.040	0.125	*0.955*	-0.127	0.052	-0.363
DIVISOR_4	-0.126	0.105	*0.977*	-0.151	0.089	-0.384
DISCRSE_1	0.410	-0.245	-0.331	0.315	0.292	**0.854**
DISCRSE_2	0.204	**0.719**	-0.271	0.158	0.139	**0.609**
DISCRSE_13	0.202	*-0.542*	-0.223	0.164	0.200	**0.573**
DISCRSE_3	**0.907**	0.180	0.065	0.292	*0.789*	0.279
DISCRSE_4	0.097	0.115	**0.939**	-0.092	0.141	-0.375
DISCRSE_5	*0.537*	**0.807**	0.142	0.079	0.419	0.031
DISCRSE_6	**0.853**	0.210	0.033	0.243	*0.517*	0.337
DISCRSE_8	**0.741**	0.117	0.071	0.194	*0.653*	-0.068
DISCRSE_9	**0.569**	0.030	0.104	0.247	*0.676*	0.114
SYNTAX_1	*0.729*	0.111	-0.169	0.182	0.413	*0.590*
SYNTAX_2	0.306	0.150	0.116	-0.166	**0.650**	0.046
SYNTAX_3	*0.718*	-0.051	-0.043	0.360	**0.553**	0.232
SYNTAX_4	0.424	0.028	0.003	0.058	**0.703**	0.237
SYNTAX_5	*0.775*	0.108	-0.020	0.237	**0.813**	0.410
CONTENT_1	0.315	-0.085	-0.190	**0.942**	0.123	0.348
CONTENT_2	0.260	0.064	-0.185	**0.834**	0.008	0.200
CONTENT_3	0.215	-0.034	-0.030	**0.936**	0.008	0.166
Characteristic	ARG.-1	ARG.-2	ARG.-3	ARG.-4	ARG.-5	ARG.-6

□ = Factor-feature bundle (major feature loadings in **bold**)

▨ = Echo bundle (major feature loadings in *italics*)

▒ = Features used as divisors only, not used in essay scoring (therefore disregarded in factor naming)

Table 3

Summary of Results for the Principal Factor Analysis:
Inter-Promax Factor Correlation Matrix for the Estimation Data
(Kappa = 25)

Issue Generic Model
(25 features)

			Factor			
Factor	1	2	3	4	5	6
1	1.000					
2	0.145	1.000				
3	0.289	0.204	1.000			
4	$\boxed{0.623}$	0.198	0.221	1.000		
5	0.462	0.049	0.028	0.447	1.000	
6	0.017	-0.204	-0.049	-0.160	0.108	1.000

▨ = Divisor feature-dominated factor
(disregarded in balance of study)

Table 4

Summary of Results for the Principal Factor Analysis:
Inter-Promax Factor Correlation Matrix for the Estimation Data
(Kappa = 25)

Argument Generic Model
(21 features)

			Factor			
Factor	1	2	3	4	5	6
1	1.000					
2	0.139	1.000				
3	-0.015	0.119	1.000			
4	0.418	-0.053	-0.158	1.000		
5	$\boxed{0.818}$	0.152	0.076	0.288	1.000	
6	0.166	0.073	-0.420	0.189	0.074	1.000

Table 5 summarizes the model-fit statistics examined for both the exploratory and confirmatory procedures. The two six-factor models appeared to fit the cross-validation samples only slightly less well than they fit their respective model-building samples. The confirmatory factor models provided information on the fits of the exact feature-to-factor structures hypothesized in the six-factor models, with marginal effects of other features removed from each factor. The results of this analysis, as anticipated, showed substantially less explanatory power, suggesting that limiting the loading of certain features to just a single factor takes away from the dynamic interplay of features (i.e., the presence of certain features in multiple factors) in an essay score.

Table 5

Summary of Results for the Principal Factor Analysis:
Model-to-Data Fit Estimation (Exploratory and Confirmatory Models)
(Kappa = 25)

		Issue		Argument	
Fit Statistic		Model	Cross-Val.	Model	Cross-Val.
Exploratory Model:					
Variance Explained	reduced feature set	73%	68%	79%	76%
	full feature set	48%	--	49%	--
Reproduced Correlation Residuals > .05	reduced feature set	23%	24%	14%	18%
	(largest)	-.28, -.20	-.26, -.23	-.19, -.15	-.28, -.18
	full feature set	16%	--	14%	--
Communalities	> .70	15/25	12/25	15/21	13/21
	> .50	22/25	19/25	20/21	20/21
Bartlett's χ^2 (df_{Issue} = 300, $df_{Argument}$ = 210		4987	7179	4615	16049
Confirmatory Model:					
Variance Explained by Factors (range)		--	22-38%	--	18-34%
Adjusted Goodness-of Fit Index		--	.78	--	.73
χ^2 (df_{Issue} = 140, $df_{Argument}$ = 213)		--	834	--	1122
RMSEA		--	.11	--	.15

Five of the six factors in the best-fitting "issue" factor model were proposed for retention as "underlying characteristics" for further study, while all six factors in the best-fitting "argument" model were proposed for retention as "underlying characteristics." Since factor analysis is a purely mathematical technique, each underlying characteristic was also carefully reviewed by several ETS experts in essay test development to assess whether it could be sensibly operationally defined for expert raters. All eleven of the underlying characteristics survived this expert "reasonableness review," although the test development experts predicted that several of the characteristics would not be seen as valid by expert raters. This proved to be the case, as will be discussed shortly. An ordered list of the underlying characteristics and their associated features, ranked by contribution to the explained variance in the *e-rater* data, is provided in Table 6.

Phase II: Reflexivity of the Task and Domain Structure: In this phase, the intent is to investigate whether the hypothesized structure of underlying writing characteristics represents a set of constructs of writing skill, as identified by experts, and not simply a series of rote counting tasks to be completed. Also, a modified *e-rater* model, consisting of a series of non-overlapping "submodels," is created for each essay type. Each submodel generates an essay score based solely on the set of *e-rater* features comprising the underlying characteristic the submodel identifies.

A set of clear, operationally defined rating criteria were drafted for the underlying characteristics, with each rating criterion precisely paralleling a single corresponding underlying characteristic from Table 6. These rating criteria were then used to create characteristic-specific rating rubrics for each of the two essay prompt types. Five characteristic-specific rubrics were created for the "issue" prompt type, and six for the "argument" prompt type. Once the characteristic-specific rubrics were completed, several experts from ETS Test Development reviewed the rubrics and provided feedback on their clarity and expected usability by expert raters. A number of modifications were suggested to improve the interpretability of the rubrics.

Table 6

Ordered List of the Underlying Characteristics and Their Associated Features
(ranked by pre-rotation contribution to explained variance in the *e-rater* data)

Issue Generic Model (25 features)		Argument Generic Model (21 features)	
Characteristic (pre/post-rotation eigenvalues)	Feature	Characteristic (pre/post-rotation eigenvalues)	Feature
ISSUE-1	DISCRSE_3	ARGUMENT-1	DISCRSE_3
(7.50 / 6.89)	DISCRSE_6	(5.79 / 5.28)	DISCRSE_6
	DISCRSE_8		DISCRSE_8
	DISCRSE_7		DISCRSE_9
ISSUE-2	DISCRSE_4	ARGUMENT-2	DISCRSE_5
(2.77 / 4.67)		(3.88 / 3.11)	DISCRSE_2
ISSUE-3	SYNTAX_4		
(1.96 / 6.36)	SYNTAX_5	ARGUMENT-3	DISCRSE_4
	SYNTAX_2	(2.64 / 3.27)	
		ARGUMENT-4	CONTENT_1
ISSUE-4	DISCRSE_1	(2.20 / 3.51)	CONTENT_3
(1.35 / 5.18)	DISCRSE_2		CONTENT_2
ISSUE-5	CONTENT_3	ARGUMENT-5	SYNTAX_5
(1.12 / 2.25)	CONTENT_1	(1.09 / 4.67)	SYNTAX_4
	CONTENT_2		SYNTAX_2
		ARGUMENT-6	DISCRSE_1
		(1.03 / 2.61)	DISCRSE_2
			DISCRSE_13

While an underlying characteristic is identified by its unique combination of associated writing features from the factor analysis, its corresponding characteristic-specific rubric must not simply be a "laundry list" of the subsumed writing features, but rather a scoring guide that, without actually identifying the features specifically, prompts the rater to focus on qualities of the writing that parallel, coincide with, or capture the features. From this perspective, experts from ETS Test Development were asked to evaluate the premise that something more than rote counting was being called for by each characteristic-specific rubric. The end product of this examination was two sets of rubrics that, while prompting a rater to look for evidence that coincides with the desired underlying features, still call for identification of writing quality that transcends rote counting of incidences.

Phase III: Substantive Theories and Process Models: In this phase, the question to be answered is, "is the study addressing the right things?" That is, the study probes whether the cognitive processes raters engage during essay rating actually reflect *both* the processes identified as the intended targets, in Phase II, *and* the processes assumed to be emulated by *e-rater*. This was accomplished by the procedures that follow.

The newly written characteristic-specific scoring rubrics were used by a group of four expert raters, along with the original holistic rubrics, on a fresh sample of GRE Writing Assessment essays taken from the cross-validation essay samples discussed earlier. Each expert rater rated a sample of 110 essays twice, using on each essay first the holistic rubric normally used to score GRE Writing Assessment essays operationally and then the characteristic-specific rubrics designed for this study. In order to expedite the process, the initial ratings were done by these raters at home. While it is generally preferable to have raters interacting on-site in a "conferencetype" atmosphere when rating, as this has been shown to have a small, positive effect on rating reliability comparative to remote-site rating (Breland & Jones, 1988), this particular group of expert raters, quite familiar with the essay types and scoring rubrics, were believed by ETS staff to be capable of producing reliable ratings regardless of their location.

Several weeks later, the group arrived at ETS, two raters on each of two consecutive days, to score their same samples of essays using the characteristic-specific rubrics. The raters were selected specifically on account of their expertise in GRE essay scoring and their previous participation in research activities with the GRE program. These four raters, known as "scoring leaders," are all currently faculty in English programs at colleges and universities in the Delaware Valley area. ETS Test Development staff have a high degree of confidence in the abilities of these raters, a confidence which proved critical as the challenging procedures that follow were carried out.

Upon their arrival at ETS on the day of their face-to-face participation in the project, each rater was provided a thorough introduction to the research and the specific activities scheduled to take place that day. The characteristic-specific scoring rubrics were then presented. This was the first time any of the raters has seen these rubrics. All procedures, except as noted, were recorded on audio tape. The first procedure was a roughly 30-minute semi-structured group interview, in which the raters were asked about their first impressions of the characteristic-specific rubrics. Specifically, they were asked the following questions about each rubric in turn:

- Do you find the rubric to be interpretable, meaningful, and useable to you? That is, do you feel, on first impression, that you will even be able to identify the characteristic that this rubric calls for you to evaluate in an essay?

- Do you find the characteristic targeted by the rubric a *justifiable* trait of an essay, worthy of being evaluated individually?

- Is it possible for you to conceive, on first impression, how you would go about distinguishing among the six possible score categories with respect to the characteristic identified in this rubric?

Almost immediately, some strong impressions emerged from the raters on rubrics for both the "issue" and "argument" essay types. For both essay types, the raters had problems rationalizing the legitimacy of looking for two of the characteristics: ISSUE-2/ARGUMENT-3 and ISSUE-4/ARGUMENT-6 (the compound label indicates that a characteristic appears in both essay types). All four raters indicated that it was unusual for them to look at such characteristics in isolation, and they all questioned the legitimacy of scoring an essay exclusively for such characteristics, even when the characteristics are to be taken as part of a larger composite scoring of the essay. Also, they found the very prospect of looking for these characteristics to be something unfamiliar to them; adjectives like "unnatural" and "discomforting" were heard in their description of the impending task they would perform.

Two other characteristics, ISSUE-1/ARGUMENT-1 and ISSUE-5/ARGUMENT-4, were perceived by all four raters as difficult to parse into separate, ratable qualities as, by their account, it was odd for them to think in terms of separating the content of the essay from the type of discourse used. In fact, it emerged later each day that the raters typically considered content to be relevant to an essay only inasmuch as it relates to the positions being taken, and that idle mention of prompt-related material would not normally constitute "content" for them. (This issue presents one example of the conceptual "retraining" that had to be undertaken, by both the researcher and the raters, in order to help the raters to see the task as the researcher construed it.)

Interestingly, the concept introduced by Muraki, Lee, and Kim (2000) and noted in the factor analysis, namely, the dynamic relationship hypothesized between syntax and discourse structures in the construction of an essay, did not manifest itself as much in the initial discussions (although there is some evidence that it emerged in a later procedure). That is, the raters did not report anticipating difficulty in separating the syntactic structure of the essay from the type of discourse used, although several raters agreed that these qualities are highly associated in some instances.

To examine in greater detail whether the expert raters were following processes that focus implicitly in some way on the features analyzed by *e-rater*, each rater was asked to "talk aloud" while rating essays. Using a procedure outlined by Ericsson and Simon (1980) as a guide, each rater was instructed to verbalize all her thoughts as she rated two essays, using each of the characteristic-specific rubrics in turn. Intermittently, she was stopped by the researcher and asked to recall what thought processes had just gone through her mind during the most recent rating. This exercise was performed initially under the observation of a resident expert in "talk-aloud" procedures at ETS. The four "talk-aloud" sessions produced several hours of audio tape, which were then converted into coded protocols, or bullet-like summary transcriptions of what was said by the rater during each session.

Several issues emerged from a review of the protocols. First, the protocols suggest that the essay rating process is highly complex and interactive with many components of the rater's cognitive processing structure, to the point that it is essentially imbedded in the processing pattern. That is, the data suggest that, even with considerable advance discussion (and, in the case of the two raters who sat for the "talk-aloud" session in the afternoon, a half-day of hands-on experience), raters may have had some difficulty adapting to the specific requirements of the characteristic-specific rubrics, tending to slip back into their more familiar holistic scoring paradigm, even as they *tried* to accommodate the characteristic-specific scoring task presented. Evidence of this includes statements suggesting reversion due to frustration, such as "oh, well, I'll give this a four," and statements suggesting uncertainty of reference: "well, I *think* this is a three."

Second, not surprisingly, the protocols suggest dramatically differing personal styles of reaching a scoring conclusion. Not only did the amount of time to produce a score vary widely across both raters and rubrics, but the engaged processes often seemed incongruous with the task presented. This is not an unusual finding in "talk-aloud" protocols, however, since the context-lending descriptions of the processes are absent; the participants had not had more than minimal prior practice and, as such, had not thoroughly processed how to "keep talking" throughout the session.

There was some evidence that raters were looking for words, phrases, transitional terms, and other indicators of discourse, syntax and content. Occasionally, there was reiteration of terms used in the rubrics, indicating that the rater was "mindful of," if not actually looking for, evidence of the characteristic of interest. Many of these indicators were recognizable as features that *e-rater* targets; others were terms that likely co-occur with one or more features. Overall, while it is largely the case that the raters were *not* actually counting occurrences of indicator cues representing *e-rater* features, they were tracing qualities that incorporate such features. The degree to which this was consistently done, however, is impossible to determine more precisely from the protocols.

Next, each rater participated in what Ericsson and Simon (1980) refer to as a "social verbalization" exercise, in which the rater scores an essay interactively with the researcher and, in the final step, with another rater as well. This exercise proved to be most useful as a training process and, in retrospect, probably would have been more beneficial to all four raters if it had been conducted at the beginning of the day. In this exercise, the raters finally "learned" what the researcher had in mind as the right way to score each characteristic, and compared this to the process they had each developed independently for scoring each characteristic up to that point. Several of the raters indicated, after this exercise, that they had significantly changed their understanding of what was being sought by one or more of the characteristics (particularly the topic-relevant content characteristic) and, therefore, should revisit some of the scorings they had done earlier in the day. Each rater that indicated the inclination to do this also felt that she could do so in a consistent manner, without biasing either her earlier scoring or the scoring yet to be done.

Each day concluded with a short debriefing session, preceded by a "normalization" period of silent, independent scoring intended to get each rater "into the groove." At this point, several of the raters mentioned the interplay between syntax and certain forms of discourse. Specifically, when an essay writer would make a certain type of assertion in the essay, the raters would expect to see the associated use of certain types of syntactic structures. The absence of such syntax in such an instance would render the assertion superficial. While essays with and without such syntactic variety were both seen, clearly the essays containing the syntactic variety associated with that type of discourse were viewed by the raters as superior.

E-rater was later run on each essay in the cross-validation samples, both in its original generic form and its subscore-models form, for each essay prompt type. The *e-rater*-to-expert rater proportions of agreement and correlations are presented in Table 7, along with estimates of Cohen's Kappa, a "chance-corrected measure of agreement" (Fleiss, 1981, p. 217) that is commonly reported in results of ratings. Additionally, the correlations of *e-rater* scores between the three individual prompt-specific models and the generic model for each essay type are presented as a note to each section of Table 7.

Discussion

Four topics are discussed here: the *implications* that the findings have on the research question asked; the *limitations* of the study arising from both gaps in theory and shortcomings of the implementations of the methods proposed; the *educational importance* of the research; and *potential extensions* of the research in the future.

Table 7

E-rater-to-Expert Rater Proportions of Agreement, Kappas, and Correlations:
"Issue" Generic Models
(sample size = 200 essays)

		Characteristic-Specific						
	Holistic	ISSUE-1	ISSUE-2	ISSUE-3	ISSUE-4	ISSUE-5	Field Score†	*E-rater*-to-Field Score
Exact agreement with both raters	.42	.43	.29	.38	.43	.32	.57	.31
Exact agreement with one rater	.04	.02	.04	.02	.02	.02	--	.34
Adjacent agreement with both raters	.40	.42	.48	.44	.40	.52	.41	.24
Subtotal: Exact and adjacent	**.86**	**.88**	**.82**	**.84**	**.85**	**.87**	**.98**	**.89**
Disagreement	.14	.12	.18	.16	.15	.13	.02	.11
Kappa	.24**	.23**	-.01	.12**	.04	-.02	.43**	.34**
Correlation	.56**	.44**	-.01	.32**	.26**	.18*	.79**	.63**
Exact and adjacent agreement: Interrater subsample (*n* = 19)	*1.00*	*.95*	*.95*	*1.00*	*1.00*	*1.00*	--	--
Correlation: Interrater subsample (*n* = 19)	*.95***	*.80***	*.66**	*.87***	*.83***	*.89***	--	--

Note: The correlations of the prompt-specific *e-rater* scores to the generic model *e-rater* scores are as follows:

Prompt 1 – Generic: .85** (*n* = 67)
Prompt 2 – Generic: .81** (*n* = 30)
Prompt 3 – Generic: .88** (*n* = 30)

† the score assigned previously during operational scoring; it is either the average of two expert raters' scores or, when the raters disagree by more than one point, the score assigned by the adjudicator.

* statistically significant at the α = .05 level.

** statistically significant at the α = .01 level.

Table 7 (continued)

E-rater-to-Expert Rater Proportions of Agreement, Kappas, and Correlations:
"Argument" Generic Models
(sample size = 200 essays)

		Characteristic-Specific							
	Holistic	ARG.-1	ARG.-2	ARG.-3	ARG.-4	ARG.-5	ARG.-6	Field Score	*E-rater*-to-Field Score
Exact agreement with both raters	.53	.32	.29	.31	.30	.43	.40	.53	.32
Exact agreement with one rater	.04	--	--	--	--	--	--	--	.32
Adjacent agreement with both raters	.36	.62	.50	.49	.59	.50	.52	.41	.24
Subtotal: Exact and adjacent	**.93**	**.94**	**.79**	**.80**	**.89**	**.93**	**.92**	**.94**	**.87**
Disagreement	.07	.06	.21	.20	.11	.07	.08	.06	.13
Kappa	.40**	.10*	.05	.00	.05	.21**	.11*	.47**	.40**
Correlation	.74**	.49**	.36**	-.08	.19**	.56**	.39**	.80**	.74**
Exact and adjacent agreement: Interrater subsample (*n* = 20)	*.95*	*.85*	*.85*	*.70*	*.95*	*.85*	*.95*	--	--
Correlation: Interrater subsample (*n* = 20)	*.81***	*.75***	*.67***	*.51**	*.78***	*.72***	*.77***	--	--
Exact and adjacent agreement: Adjudication subsample (*n* = 12)	--	--	--	--	--	--	--	--	*.92*
Correlation: Adjudication subsample (*n* = 12)	--	--	--	--	--	--	--	--	*.70**

Note: The correlations of the prompt-specific *e-rater* scores to the generic model *e-rater* scores are as follows:

Prompt 1 – Generic: .77** (*n* = 90)
Prompt 2 – Generic: .77** (*n* = 124)
Prompt 3 – Generic: .72** (*n* = 100)

* statistically significant at the α = .05 level.

** statistically significant at the α = .01 level.

Implications of the Findings: This investigation sought evidence for *e-rater* score "intrinsic" validity from three sources: the underlying factor structure of the *e-rater* scoring models; the degree of correlation between expert rater scores and *e-rater* scores, both holistically and on specific characteristics of writing that both expert raters and *e-rater* attend to in an essay; and the nature of the testimonials provided by expert raters as they examine essays, holistically as well as by characteristic. The first of these sources, the underlying factor structure of the *e-rater* scoring models, provided some evidence that *e-rater* counts features that do in fact signal the presence of desirable writing characteristics, such as topic-relevant content, syntactic variety, and skillful use of discourse in an essay. Evidence of this lies in the close parallel between the way that ETS Test Development experts and the GRE Writing Assessment holistic scoring guides define these characteristics and the way the corresponding *e-rater* features group in order under analogous characteristics. That is, the evidence gathered supports the statement by ETS Technologies, Inc. that *e-rater* is designed to reflect the writing qualities specified in the GRE Writing Assessment holistic scoring guides. Obviously, *e-rater* does not *read* an essay, so it cannot "look for" or "evaluate" writing qualities. However, *e-rater* can, and does in some instances, detect evidentiary traces, the proverbial "breadcrumbs in the path," that signal these qualities, using its own version of the characteristics.

The patterns of score proportions of agreement and correlations, as reported in Table 7, are a second source of evidence for *e-rater* score "intrinsic" validity. The factor analysis results had implied that the first factor in each model, named characteristic ISSUE-1/ARGUMENT-1, having the highest feature loadings, on average, and explaining by far the largest percentage of variance in the data (before rotation), would drive the *e-rater* score for either essay type. This appears to be borne out in Table 7; no other characteristic produced scores for either essay type that agree as often or correlate as highly between expert raters and *e-rater*. The syntactic variety characteristic, ISSUE-3/ARGUMENT-5, produced the second-highest correlation of all for both essay types and the highest kappa of all the "argument" characteristics. However, the topic-relevant content characteristic, ISSUE-5/ARGUMENT-4, did not fare as well, even though the researcher had presumed beforehand that identifying on-topic content in an essay would be a fairly straightforward task. The difficulties several raters cited in dissociating essay *content* from the rater's interpretation of the writer's *intent*, as mentioned earlier, might be to blame for this.

The transcripts of raters' recorded conversations with the researcher provide the most compelling evidence both for and against the "intrinsic" validity of *e-rater* scores; that is, their construct relevance and content representativeness, their reflection of the task and domain structures engaged in scoring an essay, and their agreement with the process models used by experts to generate essay scores. Specifically, all four raters agreed that the syntactic variety characteristic, ISSUE-3/ARGUMENT-5, and the topic-relevant content characteristic, ISSUE-5/ARGUMENT-4, are relevant, identifiable in an essay, reflective of what a rater should look for in an essay, and either explicitly or implicitly parts of their processing schemas when rating an essay. To a lesser extent, the raters viewed the principal discourse characteristic, ISSUE-1/ARGUMENT-1, the same way, although they had reservations about the efficacy of assessing this characteristic in isolation from the other qualities of the essay. Conversely, all of the raters viewed two of the characteristics, ISSUE-2/ARGUMENT-3 and ISSUE-4/ARGUMENT-6, as being inappropriate with respect to any aspect of "intrinsic" score validity.

Taken together, the three sources present a "mixed bag" of evidence for and against the "intrinsic" validity of *e-rater* essay scores. The strongest evidence for "intrinsic" validity comes from the same source as the strongest evidence against it: the raters themselves. From the raters' perspective, simpler scoring models that leave out certain features of the current *e-rater* models would likely be substantially more likely to be accepted by expert raters, and particularly those who are college faculty in the liberal arts and sciences or the humanities, high school English teachers, or others in related professions. The statistics arising from the present study suggest that simpler *e-rater* models might do just as well at agreeing with the scores of expert raters, although the proportion of total variance in the expert rater scores explained by the *e-rater* scores might decrease from an already modest level.

Limitations of the Study: The first limitation encountered was the limited amount of data available for the study. Since its inception in October, 1999, the GRE Writing Assessment examinee population has grown slowly, so the number of essays available (for a selected set of prompts) for study purposes was relatively small. While the sample sizes available for model building and cross-validation meet the minimums suggested by Stevens (1996), larger samples might have improved the agreement between the initial and cross-validation models.

The relatively low model-to-data fits suggest that the explanatory power of the hypothesized factor structures does not go far in helping to answer the score validity question. The researcher postulates that this is largely due to what Muraki, et al. (2000) found and what the correlations reported in the present study suggest, namely, that there is such a high level of dynamic interaction between writing characteristics that any attempt to isolate them, essentially taking them out of their natural context, produces a sort of "reverse synergy," yielding parts that are far less useful individually than they are together. This may well be an insurmountable limitation to this kind of investigation.

Another limitation to the study arises from the lack of adequate training and practice time the four raters received prior to performing the characteristic-specific scoring tasks. Optimally, raters involved in a study of this kind would be introduced to the tasks and materials well in advance and given ample opportunity to practice on exercise essays and to ask questions. Unfortunately, in the present study, the training and learning were going on as the actual study samples were scored. Fortunately, ETS Test Development staff took care to select experienced raters who were up to the tasks. The high level of expertise and adaptability of these raters likely mitigated the task-specific inexperience limitation posed here.

Still another limitation of the present study is that the characteristic-specific scoring rubrics developed for this study, while modeled after existing GRE Writing Assessment holistic scoring rubrics, were not adequately pre-tested before being used in the study. Again, the limitation is partially mitigated by the high level of expertise employed in the development of the rubrics, which included the active participation of two leading essay test development professionals. However, the tentativeness expressed by these experts with respect to some of the rubrics, combined with the unforeseen need to make considerable revisions to the rubrics following the first day of the face-to-face portion of the project, lead the researcher to believe that there may have been nontrivial inaccuracies in the rubrics, resulting in inconsistent application of the rubrics across raters. While the various discussion exercises described earlier were intended partly to address this problem, it is still likely that inconsistencies in using the rubrics contributed substantially to the low correlations produced in the characteristic-specific scorings.

Educational Importance of the Research: Commencing with the work of Carlson and Ward (1988), ETS has for many years pursued the development of computer-based free-response scoring tools. As reducing by half the number of expert rater scorings required may produce considerable cost savings for essay testing programs, the appeal of these tools is certain to grow, and perhaps even extend beyond testing organizations. The present study represents a first attempt at establishing a protocol for demonstrating the "intrinsic" validity of computer-assisted essay scores. (The term "computer-assisted" reflects the current practice with the GMAT of using *e-rater* strictly as a reliability check for one expert rater score; when that one expert rater and *e-rater* disagree, a second expert rater is called upon to adjudicate the discrepancy. Scoring of the GRE Writing Assessment is not computer-assisted.)

Potential Extensions of the Research: As the pool of available GRE Writing Assessment essays grows, a more comprehensive follow-up effort, addressing all or most of the limitations cited earlier, could produce results that confirm and significantly augment the findings presented in the present study. Also, it will become possible to investigate the behavior of expert rater and *e-rater* scores under specific anomalous circumstances, such as with "1" and "6" essays alone, with adjudicated essays, and with contrasting samples of operational essays and essays commissioned especially for a study. Separately, additional research into parallels between the cognitive processing of expert raters and the electronic processing of *e-rater* may provide further insight to strategies for improving *e-rater* from an "intrinsic" score validity perspective.

References

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13* (1), 1-18.

Breland, H. K. (1983). *The direct assessment of writing skill: A measurement review* (Research Report 83-32). Princeton, NJ: Educational Testing Service.

Breland, H. K., & Jones, R. J. (1988). *Remote scoring of essays* (Research Report 88-04). Princeton, NJ: Educational Testing Service.

Burstein, J. C., & Chodorow, M. (1999, June). *Automated essay scoring for nonnative English speakers.* Paper presented at the Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

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

Burstein, J. C., Kukich, K., Wolff, S. E., Lu, C., & Chodorow, M. (1998b, August). *Enriching automated scoring using discourse marking*. Paper presented at the Workshop on Discourse Relations and Discourse Marking at the annual meeting of the Association of Computational Linguistics, Montréal, Canada.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carlson, S. B., & Ward, W. C. (1988). *A new look at formulating hypothesis items* (GRE Board Professional Report 85-14P). Princeton, NJ: Educational Testing Service.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton-Mifflin.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Ericsson, K. A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Boston: MIT Press.

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

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods for cumulating research findings across studies.* Newbury Park, CA: Sage.

Kaplan, R. M., Wolff, S. E., Burstein, J. C., Lu, C., Rock, D. A., & Kaplan, B. A. (1998). *Scoring essays automatically using surface features* (ETS Research Report 98-39). Princeton, NJ: Educational Testing Service.

Keith, T. Z. (1999, April). *Newest evidence for construct validity in PEG.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Canada.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16,* 294-304.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.

Messick, S. (1987). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Mahwah, NJ: Erlbaum.

Messick, S. (1989b). Validity. In Linn, R. (Ed.), *Educational measurement* (3rd ed.).. Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50* (9), 741-749.

Messick, S. (1999). Technology and the future of higher education assessment. In Messick, S. (Ed.), *Assessment in higher education: Issues of access, student development, and public policy* (pp. 245-254). Mahwah, NJ: Erlbaum.

Muraki, E., Lee, Y., & Kim, R. M. (2000). *Factor analysis of NAEP e-rater data.* Unpublished manuscript.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62* (2), 127-142.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76* (7), 561-565.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997, March). *Computer analysis of student essays: Finding trait differences in student profile.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich K. (2000). *Comparing the validity of automated and human essay scoring* (GRE Board Research Report 98-08aR). Princeton, NJ: Educational Testing Service.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (1999, April). *Trait ratings for automated essay grading.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Canada.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40,* 371-396.

Stevens, J. (1996). *Applied multivariate statistics for the Social Sciences* (3rd Ed.). Mahwah, NJ: Erlbaum.

AERA

ERIC®

TM033472

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Computerized Scoring of Essays for Analytical Writing Assessment	
Author(s): P. Adam Kelly	
Corporate Source: Florida State University	Publication Date: April 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: P. Adam Kelly	Printed Name/Position/Title: P. Adam Kelly	
Organization/Address: 257 John Knox Road #L-201 Tallahassee, FL.32303	Telephone: 850-644-4952	FAX:
	E-Mail Address: pkelly@garnet.acns.fsu.edu	Date: 11/01/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com