ISSN 1552-8219

Research Reports

RR - 81
January 2008

Analytic Scoring of TOEFL CBT Essays: Scores From Humans and *E-rater*®

Yong-Won Lee
Claudia Gentile
Robert Kantor

Analytic Scoring of TOEFL® CBT Essays:
Scores From Humans and ***E-rater***®

Yong-Won Lee, Claudia Gentile, and Robert Kantor[1]
ETS, Princeton, NJ

RR-08-01

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Abstract

The main purpose of the study was to investigate the distinctness and reliability of analytic (or multitrait) rating dimensions and their relationships to holistic scores and *e-rater*® essay feature variables in the context of the TOEFL® computer-based test (CBT) writing assessment. Data analyzed in the study were analytic and holistic essay scores provided by human raters and essay feature variable scores computed by *e-rater* (version 2.0) for two TOEFL CBT writing prompts. It was found that (a) all of the six analytic scores were not only correlated among themselves but also correlated with the holistic scores, (b) high correlations obtained among holistic and analytic scores were largely attributable to the impact of essay length on both analytic and holistic scoring, (c) there may be some potential for profile scoring based on analytic scores, and (d) some strong associations were confirmed between several *e-rater* variables and analytic ratings. Implications are discussed for improving the analytic scoring of essays, validating automated scores, and refining *e-rater* essay feature variables.

Key words: Analytic scoring, multitrait scoring, holistic scoring, text features, profile scores, writing feedback, automated essay scoring, *e-rater*, ESL writing assessment, TOEFL CBT

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2007-2008) members of the TOEFL Committee of Examiners are:

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

The study was funded by the TOEFL program and partially supported by the ETS Summer Program in Research for Graduate Students Internship award to Chen-Ling Chou. Several members of the ETS staff and external reviewers in addition to the authors contributed to this project. Chi Lu provided us with *e-rater* data containing 12 *e-rater* essay feature variables for each of the essays used in this study. Nan Kong extracted the *e-rater* essay feature scores for each essay from the *e-rater* output data, and Fred Cline prepared rating data files containing analytic scores for our analyses. Regina Mercadante and Venus Mifsud provided logistical and administrative support for rater recruitment and training and actual rating of essays. Chen-Ling Chou contributed to the project by creating rating designs and rating kits for the essays. Thanks go to James Daubs for allowing us to recruit raters from the ELA (English Language Assessment) Item Writer Summer Institute held at ETS in the summer of 2003 and also to all the raters who participated in the rater training workshops and subsequent scoring of essays. Finally, we would like to thank Jill Burstein, Alister Cumming, Dan Eignor, Don Powers, Mary Enright, and Sara Weigle for their helpful feedback on earlier versions of the report. The responsibility for any errors that remain is solely ours, and the ideas and opinions expressed in the paper are those of the authors, not necessarily of ETS or the TOEFL program.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Introduction

Holistic (that is, global or impressionistic) scoring has been widely used in many large scale writing assessments, including the computer-based Test of English as a Foreign Language™ (TOEFL®), Graduate Record Examinations® (GRE®), and GMAT (Graduate Management Admission Test; Williamson, 1993; Williamson & Huot, 1993). For holistic scoring rubrics, elaborate score descriptors are usually developed for several score levels, and the writing qualities of an essay are usually represented by a single, overall *holistic* rating. One drawback of holistic scoring has to do with its inability to capture examinees' specific weaknesses and strengths in writing (Wiegle, 2002). This failure can be especially true for second language learners who are still developing their writing skills and who are thus likely to show uneven profiles across different aspects of writing. For examinees with such nonuniform patterns of proficiencies across different aspects of writing skills, analytic (or multitrait) scoring rubrics can be useful in capturing their weaknesses and strengths (Bacha, 2001; Connor-Linton, 1995; Hamp-Lyons, 1995, 1991; Raimes, 1990; Sasaki & Hirose, 1999).[2] For this reason, many educators believe that analytic scoring can be useful for generating diagnostic feedback to guide instruction and learning (Hamp-Lyons, 1995, 1991; Roid, 1994; Swartz et al., 1999).

Despite such advantages, analytic scoring has not been widely used for large scale writing assessments for several important reasons. One reason has to do with the cost associated with human rating of essays (Huot, 1990; Veal & Hudson, 1983). Even when holistic scoring is used, the scoring of writing samples poses a cost challenge for testing programs, compared to machine-scored multiple-choice items. Because analytical scoring requires multiple ratings of each essay by human raters, the number of raters and time required for rater training and scoring is much greater for analytic than for holistic scoring. In addition, analytic ratings have often proven less useful than expected because rating dimensions are often highly correlated among themselves and with holistic scores, thus rendering them redundant from a psychometric point of view (Bacha, 2001; Freedman, 1984; Huot, 1990; Veal & Hudson, 1983).

Recently, however, analytic scoring has received renewed attention in writing assessment, especially in the context of automated essay scoring and evaluation. The *e-rater*® system is one such system that is in operational use for some large scale assessment programs. *E-rater* has also been embedded in online writing practice services, such as *Criterion*^{SM}, to score essays written and submitted by students or prospective test takers (http://www.ets.org/criterion).

Although most automated scoring research has focused on emulating human holistic scores so far, in the study described here we explore the use of automated scoring (via *e-rater*) for generating analytic scores. One exciting implication of such technology is that the large rating cost traditionally associated with analytic scoring can be reduced significantly if valid analytic scores can be computed automatically. Besides, if computer-generated holistic and analytic scores can also be traced back to more microlevel essay text features, these features can be used to provide performance feedback to learners. Making these associations will help not only to make the rating process more transparent in automated essay scoring but also to specify appropriate strategies for validating automated essay scores. One approach, for instance, might entail a rigorous examination by writing experts of the essay text features used by *e-rater* (in terms of validity and appropriateness of these features).

Recent versions of *Criterion* include two main automated evaluation components: (a) *e-rater* and (b) *Critique*™ Writing Analysis Tools (*Critique* henceforth). *E-rater* provides an instant holistic score for an essay to be evaluated, while *Critique* is text analysis software that detects various grammar, usage, mechanics, and style (GUMS) errors in the essay. *Critique* flags the parts of essays that are suspected of containing such errors and directs students to relevant writing advisories available online. These errors are aggregated into four major categories of writing errors (i.e., GUMS) and these four aggregate values are transformed into linguistic accuracy ratio variables. Each of these four variables is then used as one of the scoring variables in recent versions of *e-rater* (version 2.0 and above). Despite such a connection between *Critique* errors and four of the text feature variables of *e-rater*, links between the automated score (provided by *e-rater*) and diagnostic feedback (provided by *Criterion*) seem to be rather weak in the current system. If such links are more strongly established, the feedback can become more useful to learners and the validity of automated scores can be further established. In this regard, an enhancement of this online service would be to establish in a principled way specific links among an automated holistic score, trait (or analytic) scores, essay feature variables, and performance/diagnostic feedback for individual examinees.

Validly establishing such links in the automated scoring system could provide many advantages in terms of score validation and interpretation. First, it would enable the use of *e-rater* scoring for the dual purposes of computing automated holistic scores and generating diagnostic feedback closely aligned with the scores. This feedback would identify, for individual examinees, those specific areas of writing that require improvement on a specific prompt (or a set of prompts). Second, such a scheme would provide a more defensible and transparent basis for interpretation of automated essay scores. From a score validity perspective, one of the central tenets of measurement is that test developers should be able to provide a clear and defensible interpretation of test scores. In one sense, the ultimate test of score validation for *e-rater* scoring may not be only the degree of agreement between the human raters and *e-rater* (Bennett, 2004; Bennett & Bejar, 1998). Rather, it may instead be how closely the process of automated scoring resembles the thinking of writing experts and whether the essay feature variables used in *e-rater* can also be used as a basis for generating useful diagnostic/instructional feedback for students and teachers (Lee & Kong, 2004).

Exploring the generation of useful automated trait scores in *e-rater*, however, requires the availability of valid human-assigned analytic scores to use as criteria or *targets* for *e-rater*. These human analytic scores can be used to examine if various microtext feature variables used in *e-rater* can be clustered or reorganized in such a meaningful way that they form a basis for computing automated trait scores (i.e., scores on organization, vocabulary, language use, mechanics) and composite scores. To obtain human analytic scores that can be used for such purposes, it is necessary to develop and evaluate analytic scoring rubrics for a particular writing assessment of interest (the TOEFL CBT Writing section in this study). A literature review of existing analytic score rubrics developed for English as a second language (ESL) writing assessments would also be helpful. Once we have analytic scoring rubrics for the test of interest and have scored essays for the test based on the rubrics, this will enable us to examine not only the usefulness of human analytic scoring in generating performance feedback but also the possibility of refining or reorganizing the *e-rater* essay feature variables for automated trait scoring.

The main purposes of the study described here, therefore, were (a) to investigate, in the context of the TOEFL computer-based test (CBT) writing assessment, whether distinct (separable) and reliable (dependable) analytic rating dimensions can be identified, (b) to identify examinees with nonuniform score profiles based on human analytic scores, and (c) examine the relationships of the analytic scores not only to holistic scores but also to *e-rater* essay feature variables. Another important consideration was to identify future avenues of research for creating a principled design framework for generating automated writing trait scores and writing feedback for ESL learners with respect to their essay responses to writing tasks.

Analytic Scoring and Diagnostic Feedback in Writing Assessment

In analytic (or multitrait) scoring, writing samples are rated on several important aspects of writing quality, rather than being assigned a single overall rating (Weigle, 2002). From the perspectives of score users, one important reason for favoring the multitrait scoring method is its usefulness in capturing ESL learners' weaknesses and strengths in writing and generating diagnostic feedback to guide instruction and learning (Hamp-Lyons, 1995). Another important reason for pursuing analytic scoring has something to do with raters' decision-making processes. In investigating the reactions of professors to ESL students' essays written under timed testing conditions, for instance, Santos (1988) found that readers were generally able to judge content and language independently. More recently, Cumming, Kantor, and Powers (2002) found that both experienced ESL and English as a second language (EFL) and English native speaker raters, while evaluating essays written for TOEFL CBT prompts and prototype writing tasks of TOEFL iBT, tended to divide their decision-making between two aspects of students' writing: (a) a focus on rhetoric and ideas and (b) a focus on language. This distinction, consistently evident in the raters' thinking processes while evaluating the essays, suggests that analytic features or multiple traits (rather than a single holistic scale) are inherent aspects of skilled assessors' approach to essay evaluation.

One of the best-known analytic rubrics in ESL is one developed by Jacobs and her colleagues (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981). In their rubrics, essays are rated on five different rating dimensions of writing quality, each having a different weight: content (30 points), organization (20 points), vocabulary (20 points), language use (25 points), and mechanics (5 points). Two additional examples of analytic scales are the Test in English for Educational Purposes (TEEP; Weir, 1990) and the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991). The TEEP framework consists of seven 4-point scales that cover four aspects of communicative effectiveness (relevance and adequacy of content, compositional organization, cohesion, and adequacy of vocabulary for purpose) and three accuracy dimensions (grammar, mechanical accuracy/punctuation, and mechanical accuracy/spelling). In contrast, the Michigan Writing Assessment framework contains three 6-point scales: ideas and arguments, rhetorical features, and language control. Analytic rubrics adopted for this study are slightly modified versions of the six rating rubrics developed by a panel of ESL writing experts for a research study conducted by Gentile and her colleagues (Gentile, Riazantseva, & Cline, 2002). The six rating scales cover five major analytic rating dimensions including development, organization, vocabulary, language use, and mechanics. In a sense, this framework is similar to the Jacob et al. (1981) five-dimension rating scheme. One noteworthy difference, however, is that the *language use* dimension is further divided into two subdimensions of sentence variety/construction and grammar/usage accuracy in the Gentile et al. analytic framework (see the Rating Procedures section and Appendix A for more details).

Some researchers argue that the intent of holistic scoring is to focus raters' attention on the strengths of writing, not on its deficiencies. Jarvis and others (Jarvis, Grant, Bikowski, & Ferris, 2003), for instance, have pointed out that the ESL learners can compensate for potential deficiencies in their writing by capitalizing on a few of their strengths. It is also possible that some of the essay features are complementary in nature in terms of their contribution to the overall quality of the essays. Another important point is that different raters can also assign the same holistic score by using somewhat different rating criteria (or weighting the same criteria somewhat differently). All of these factors can potentially complicate the interpretation of holistic scores. In this respect, analytical scoring rubrics are generally known to provide more useful diagnostic feedback about examinees' writing skills (Bacha, 2001; Cohen, 1994; Hamp-Lyons, 1991; Jacobs et al., 1981; Kondo-Brown, 2002; Wiegle, 2002). This can be particularly true for second language learners who may have uneven profiles of performance across different aspects of writing (Weigle, 2002).

In terms of the relationship between analytic and holistic scores, one interesting issue is whether content and development dominate holistic judgments and thus most holistic scores can be assumed to focus on such dimensions (Cumming et al., 2002). This issue has important implications for instructional practices associated with writing feedback for ESL learners in multiple-draft writing contexts. It is often recommended that feedback on content be provided on early drafts, whereas feedback on form be provided only for later drafts (Ashwell, 2000; Silva & Brice, 2004). However, recent studies suggest that second language learners prefer to receive feedback on form and content simultaneously on the same draft (Ashwell, 2000; Fathmann & Whalley, 1990; Ferris, 1995, 2002).

As mentioned previously, one argument against analytic scoring is that analytic scores from different rating dimensions are often highly correlated among themselves and with holistic scores and thus redundant (Bacha, 2001; Freedman, 1984; Huot, 1990; Veal & Hudson, 1983). This implies that most of the learners will have uniform score profiles in relation to their performance on a prompt (or across a set of prompts), and these learners' score profiles can be represented more parsimoniously by a single overall proficiency (holistic) score. Nevertheless, as long as there are subgroups of the learners whose score profiles are deviating from a uniform score pattern each for the rest of the group, one can argue that analytic scores can still be useful in identifying areas of writing requiring further improvement for such groups of learners.

Models of Effective Linking of Components of Automated Essay Evaluation

One important motivation for undertaking this project was to explore the possibility of using the *e-rater* essay feature variables for the dual purposes of computing automated (holistic/trait) essay scores and generating performance feedback for ESL learners in the context of Web-based writing practice services. Table 1 classifies the major approaches to automated essay evaluation into four different types of systems for comparison (Lee & Kong, 2004). To outline an alternative approach to essay evaluation at a conceptual level, representation of each approach was attempted in terms of (a) the main aspects of essay evaluation under that approach and (b) the relationships among these different evaluation components in a larger system of writing assessment and instruction.

Approaches 1 through 3 were adopted for essay evaluation for an earlier version of *e-rater* (version 1.3). First, Approach 1 represents an approach in which a large number of text feature variables (so-called proxy features) are used in the scoring-model building process to predict holistic human scores, and a much smaller subset of these features is selected through a statistical procedure to score essays for a particular prompt (Burstein et al., 1998; Sheehan, 2003). Therefore, the number and types of variables selected for the scoring model can vary from prompt to prompt. In such an approach, scoring accuracy is a primary concern, whereas the validity and usefulness of individual feature variables employed may not be an important concern. Second, Approach 2 represents an approach in which an automated tool reviews students' essays, detect errors in their essays, and guide students in the essay-revision process, as in *Critique* (Leacock & Chodorow, 2003) or *Writer's Workbench* (McDonald, Frase, Gingricj, & Keenan, 1982; Reid, 1986). One of the critical functions of such systems is to flag a word or a string of words suspected of containing writing errors and to bring the learner's attention to them. One of the advantages of such an automated system is that immediate feedback can be provided about writing errors, but scoring an essay is not an immediate concern. Finally, Approach 3 represents a situation in which these two components are combined into a single system to provide both automated scores and feedback to students, as in early versions of *Criterion*. In this approach, however, the automated scores and writing feedback are generated either from two separate, disconnected modules residing in the same system or from a common natural language processing (NLP) engine, with automated scores and individual writing feedback loosely (or partially) linked to each other.

Table 1

Four Different Approaches to Automated Essay Evaluation

	Current approaches			Alternative approach
	Approach 1	Approach 2	Approach 3	
Purposes	Scoring (or classification)	Feedback (diagnostic/ formative)	Scoring & feedback	Feedback linked with scoring
Validity evidence	Rating accuracy (reliability)	Usefulness (impact)	Accuracy & usefulness	Accuracy, usefulness, & substantiveness
Variable use	Parsimony	Maximization	Parsimony & maximization separately	Maximization through hierarchical structuring
Products	Earlier versions of *e-rater*	*Critique*	*Criterion*	New generations of *Criterion*

A more ideal, alternative approach envisioned here is a system in which the set of essay features used to compute the essay scores are also used to generate useful writing feedback to the students. This means that, in the alternative automated evaluation system, the automated holistic (or composite) essay scores are linked to meaningful, automated analytic scores, which are then also linked to individual essay features and writing feedback. In addition, the individual essay feature variables should be linked to the rating criteria implied in the scoring rubrics for individual testing programs. This may be done in the sprit of making the scoring variables (or algorithms) more transparently related to both the rating criteria and theoretical components of writing proficiencies (Chung & Baker, 2003) and emulating the essential features of effective grading and feedback as accomplished by expert writing teachers in instructional settings. In discussing the different models of automated essay evaluation, it is also useful to distinguish between timed and untimed assessment contexts. Most of the large scale writing assessments are intended for admission purposes and timed-essay writing schemes are often used in such contexts. For TOEFL CBT writing, for instance, examinees are given 30 minutes to complete an essay. In such a scheme writing fluency (or productivity) often becomes a predominant dimension for holistic judgment of essay quality. Approach 1 can be effective for such assessment contexts. In contrast, untimed, multiple-draft writing schemes are most often used in instructional settings. Approach 2 in Table 1 seems to be more appropriate for untimed, multiple-draft writing situations including test preparation and instructional settings. Approach 3 and the alternative approach in Table 1 can probably be used both for large scale assessments and instructional settings.

E-rater and Essay Feature Variables

Earlier versions of *e-rater* used more than 50 text feature variables in the building of scoring models. A much smaller subset of these features (about 8–12 features) is usually selected through a stepwise regression procedure to score essays for a particular prompt (Attali & Burstein, 2006; Burstein et al., 1998; Monaghan & Bridgeman, 2005; Sheehan, 2003). For this reason, the number and kind of features selected for the scoring model vary from prompt to prompt to some extent. In contrast, more recent versions of *e-rater* (version 2.0 and above) use a more standardized set of 12 essay features to score essays across prompts (see Attali & Burstein, 2006, for more details). An attempt is made in this study to examine the relationship between human analytic scores and automated essay feature variables used in a recent version of *e-rater*. Table 2 shows a list of 12 essay feature variables used in *e-rater* version 2.0. These include two organization/development features, four linguistic accuracy ratio variables, three lexical sophistication variables, two prompt-specific vocabulary usage variables, and one essay length variable. Some of these feature sets are derived from microlevel text features that are used to provide writing feedback to the students who are writing essays in *Criterion* (e.g., two organization/development features, four linguistic accuracy ratio variables). It should also be noted that all of the four linguistic accuracy ratio variables were transformed variables from the original error ratio variables. For each of the linguistic accuracy variables, the accuracy ratio variable was one minus the original error ratio variable. In addition, the log-transformed values for these accuracy variables are computed and sometimes used in *e-rater* (see the Data Source section and Appendix C for more details).

Table 2

Definitions of 12 Essay Feature Variables Used in E-rater

Text feature category	Feature variable	Definition
Organization & development	Discourse unit score	Difference between the actual and optimal number of discourse units.
	Length of discourse unit	Average length of discourse units.
Lexical sophistication	Type/token ratio	Ratio of types (unique words) over the total number of words among content words.
	Word length	Average number of characters across words.
	Vocabulary level	Vocabulary level in terms of word frequency.
Prompt-specific vocabulary usage	Word-vector score	The essay score point value (1–6) for which the maximum cosine correlation over the six score point correlations.
	Word-vector correlation	A cosine correlation value between the essay vocabulary and the sample essays at the highest essay score point (6).
Linguistic accuracy	Grammatical accuracy ratio	1 – (number of grammar errors ÷ total number of words)
	Usage accuracy ratio	1 – (number of usage errors ÷ total number of words).
	Mechanical accuracy ratio	1 – (number of mechanical errors ÷ total number of words)
	Stylistic accuracy ratio	1 – (number of style errors ÷ total number of words)
Essay length	Total number of words	The total number of words in each essay

Research Questions

As previously discussed, one important drawback of multitrait scoring suggested by some researchers is that analytic scores for different traits often turn out to be highly correlated, not only among themselves but also to the holistic score. One related concern for the analytic scoring is that some raters can unconsciously fall back on holistic methods while doing analytic scoring (Bacha, 2001; Weigle, 2002). For this reason, it is important to provide enough scoring guidelines to the raters for each of the rating dimensions. Besides, in investigating the distinctiveness of analytic rating dimensions, it is also important to find and use appropriate, advanced statistical methods that would allow for more in-depth and rigorous analysis of the analytic scores than simple comparison of correlations. Another important goal of this study was to explore ways to refine the existing *e-rater* essay feature variables and identify new *e-rater* feature variables that should be created for the purpose of automated trait scoring and feedback in the context of ESL writing assessment. To this end, an attempt is made in this study to investigate the relationships between the analytic score assigned by human raters and the existing *e-rater* essay feature variables.

More specifically, the current program of research was conducted with the following four questions in mind:

1. How dependable are the analytic ratings assigned by human raters?
2. What are the relationships between holistic essay score and various analytic scores and the relationships among the analytic scores?
3. Can nonuniform score profiles be identifiable across six rating dimensions?
4. What essay feature variables are most closely related to each of the six different analytic dimensions for a recent version of *e-rater*?

Methods

Data Source

Data analyzed included 1 holistic essay score, 6 analytic essay scores, and 12 *e-rater* (version 2.0) essay feature variable scores obtained for a total of 930 essays written by 930 examinees for two TOEFL CBT writing prompts (see Table 2). The writing section of TOEFL CBT consists of a single essay prompt that is selected for each examinee from a pool of prompts. In this study, half of the examinees took one prompt, and another half took the other prompt. Table 3 shows prompt number and prompt content, and sample sizes for the two TOEFL CBT writing prompts used in this study: (a) history/literature vs. science/math and (b) practicing sports.

Table 3

TOEFL CBT Writing Prompts Used in the Study

Prompt ID	Prompt topic	Sample size
Prompt 1	Do you agree or disagree with the following statement? It is more important for students to study history and literature than it is for them to study science and mathematics. Use specific reasons and examples to support your opinion.	465
Prompt 2	Some young children spend a great amount of their time practicing sports. Discuss the advantages and disadvantages of this. Use specific reasons and examples to support your answer.	465

For each of these two prompts, a sample of 465 essays was selected from a larger pool of essays and used for analytic scoring, as well as for textual analysis by a recent version of *e-rater* (version 2.0). To create this sample, two separate, smaller data sets were combined and used for analytic rating for each prompt: (a) a stratified sample of 265 essays and (b) a random sample of 200 essays. Since the lowest score point of 1 is rarely used by raters for TOEFL CBT essays, it is often difficult to represent this score category in a random sample. For this reason, the stratified sampling was done for the first sample to cover a full range of essay scores, including the lowest score point of 1. The stratified sample consisted of 50 essays for each of the score categories from 2 to 6 and 15 essays for the score category of 1 to represent the entire holistic score range.

Table 4 shows the means and standard deviations of the holistic and analytic scores used in this study. The holistic scores were obtained from the operational TOEFL CBT data, and were based on two independent readings and holistic ratings of the essay response on a 1 to 6 scale. In most of the analyses, the average of the two independent ratings was used, which ranged from 1 to 6 with possible scores in intervals of 0.5. The holistic score used in this study was the average of the first two ratings before adjudication. In contrast, analytic scores were obtained for each essay by rerating the essays on six different analytic rating dimensions, which included development (DEV), organization (ORG), vocabulary (VOC), sentence variety/construction (SVC), grammar/usage (GU), and mechanics (MEC). The development and organization scores were on scales of 1 to 6, whereas the rest of the analytic scores were on scales of 1 to 5. Each of these six analytic scores used in this study was the average of two independent ratings for each essay, which ranged from 1 to 6 for the first two dimensions and from 1 to 5 for the remaining four dimensions, with possible scores in intervals of 0.5.

Table 4

Means and Standard Deviations for the Holistic and Analytic Essay Scores Used in the Study

Essay scores		Score range	Prompt 1		Prompt 2	
			M	SD	M	SD
Holistic		1–6	3.6	1.4	3.6	1.4
Analytic	DEV	1–6	3.6	1.2	3.6	1.1
	ORG	1–6	3.8	1.2	3.9	1.1
	VOC	1–5	3.1	1.2	3.0	1.2
	SVC	1–5	3.1	1.3	3.1	1.3
	GU	1–5	2.8	1.2	2.8	1.2
	MEC	1–5	3.1	1.2	3.3	1.2

Note. n = 465. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics.

Table 5 shows the means and standard deviations of the 12 *e-rater* essay feature variables used in this study. These include two discourse and organization variables (discourse unit score, length of discourse unit), three lexical complexity variables (type/token ratio, word length, vocabulary level), two prompt-specific vocabulary usage variables (word-vector score, word-vector correlation), four GUMS (grammar, usage, mechanics, and style) accuracy ratio variables, and one essay length variable (total number of words in an essay). These 12 essay feature variable scores were computed by a recent version of *e-rater* (version 2.0) for each of the 930 essays used in this study.

Rating Procedures

Analytical rating scales developed for the Gentile et al. (2002) study were modified and used for this study. In the Gentile et al. study, a panel of three ESL writing experts identified dimensions of effective essay writing based on a critical analysis of the holistic scoring rubrics for TOEFL CBT and the Test of Written English™ (TWE®), results of the pilot study, and examinee essay samples. Through reading, critical analysis, and discussion, the team identified six rating dimensions as central to effective essay writing; these included development, organization, vocabulary, sentence variety/construction, grammar/usage, and mechanics. Each of these six dimensions is described in detail in Appendix A. For the first two rating dimensions (development, organization), the examinee essays were scored on a scale of 1 to 6, as was done for the holistic rating, but for the remaining four dimensions, the essays were scored on scales of 1–5. The same rating rubrics and designs were used for the two prompts.

Please also note that log-transformed values were also computed for the four accuracy ratio variables and are reported in parentheses in Table 5. Since some of these ratio variables often turn out to have extremely small variances, the log-transformed variables are sometimes used in recent versions of *e-rater*, instead of the ratio variables.

Table 5

Means and Standard Deviations for the 12 E-rater Variables Used in the Study

Variable name	Prompt 1		Prompt 2	
	M	SD	M	SD
1. Discourse unit score	–3.40	2.42	–3.34	2.24
2. Length of discourse unit	42.50	24.20	39.80	18.90
3. Type/token ratio	0.36	0.11	0.35	0.10
4. Word length	4.70	0.33	4.60	0.29
5. Vocabulary level	52.70	6.60	56.00	5.90
6. Word-vector score	4.82	1.18	4.30	1.29
7. Word-vector correlation	0.19	0.06	0.19	0.07
8. Grammatical accuracy ratio (log)	0.99 (4.56)	0.01 (0.80)	0.99 (4.52)	0.01 (0.81)
9. Usage accuracy ratio (log)	1.00 (4.91)	0.00 (0.69)	1.00 (4.95)	0.00 (0.67)
10. Mechanical accuracy ratio (log)	0.96 (2.58)	0.04 (1.29)	0.96 (2.29)	0.04 (1.07)
11. Stylistic accuracy ratio (log)	0.88 (3.34)	0.11 (0.83)	0.86 (3.43)	0.10 (0.83)
12. Essay length	207.70	103.50	214.50	105.50

Note. n = 465.

Raters were recruited from two different pools of ESL teaching practitioners: (a) participants in the ELA (English Language Assessment) summer institute on item writing held at ETS in the summer of 2003 and (b) trained Online Scoring Network (OSN) essay raters for TOEFL CBT. Two separate, full-day training sessions were conducted, one for development and organization dimensions and the other for the remaining four dimensions (vocabulary, sentence variety/construction, grammar/usage, and mechanics). There were three rater groups, one group that scored only on development and organization dimensions; another group that scored only on the vocabulary, language use (sentence variety/construction, grammar/usage), and mechanic dimensions; and a third group that scored essays across all of the six rating dimensions for only a subset of essays. The raters in Group 3 attended both of the two rater training sessions. For actual scoring of essays, online scoring kits were prepared so that raters could rate the essays on a computer at a place they chose.

For each prompt, 400 essays were rated by two independent raters selected from a pool of 15 trained raters from Rater Groups 1 and 2. The remaining 65 essays were rated on all of the six dimensions by the three raters from Rater Group 3 according to a crossed design (person × rater × dimension). For Rater Group 3, we computed and compared the overall rating agreement rates between a pair of raters for all of the three possible pairs (i.e., Raters A and B, Raters A and C, Raters B and C). We selected a pair of raters whose overall rating agreement across essays and prompts was higher than other possible pairs of two raters, and these final two raters' ratings were averaged for each of the essays and used in the final analysis.

Data Analysis

Several statistical methods were used to analyze both the analytic and holistic scores obtained for the TOEFL essays used in this study. These analyses included (a) generalizability theory (G-theory), (b) zero-order and partial correlations, (c) multidimensional scaling (MDS), (d) Rasch item response theory (IRT), and (e) cluster analyses. G-theory analyses were conducted to examine the dependability (reliability) of the analytic scores for TOEFL CBT essays. Both correlation and MDS analyses were conducted to examine the empirical relationships among the holistic and analytic scores. In particular, MDS analysis (Borg & Groenen, 1997) was used to obtain a graphical representation of the structural relationships among the holistic and analytic scoring dimensions (especially in terms of the degree of dissimilarity). Rasch IRT analyses were conducted on the six analytic scores to identify examinees with nonuniform score profiles. Cluster analysis was done only on the subgroup of examinees (or essays) with nonuniform score profiles (who were identified based on relatively large Rasch IRT misfit statistics). Finally, correlations were also computed between the holistic and analytic scores and the *e-rater* essay feature variables. More detailed descriptions of each of these analyses follow:

G-theory analyses. Multivariate G-theory analyses were conducted on the six analytic scores obtained for this study using the computer program mGENOVA (Brennan, 1999). In the multivariate analyses, the six rating dimensions were treated as a fixed facet. A separate analysis was conducted for each of the two prompts. In each analysis, the single-facet crossed design (p• × r′•) with persons (p) as the object of measurement and with ratings (r′—first and second ratings) as random facets was employed in the G-study to estimate the variance components for each rating dimension and the covariance components for such dimensions in the G-study.[3] It was assumed that persons (p) and ratings (r′) were crossed with the rating dimensions (*v*).

Analysis of correlations among holistic and analytic scores. Correlation matrices for the holistic and analytic scores were obtained to examine the relationships among these scores for the two prompts. Two different types of correlations were computed: (a) Pearson (zero-order) correlations among holistic and analytic scores and (b) partial correlations computed after partialing out the impact of essay length on the scores. Both the total number of words (TNW) and TNW-squared values were used as covariates in computing the partial correlation to control the linear and quadratic effects of essay length on the correlation.

Multidimensional scaling analyses. MDS analyses were conducted for each of the prompts using the computer program SPSS version 12. MDS is a statistical method that represents similarity (dissimilarity) data among pairs of objects, items, or dimensions as distances among points in a low-dimensional, geometric space (Borg & Groenen, 1997). The similarity (dissimilarity) measures can be intercorrelations or ratings of similarity between the objects. In this study, the original averaged ratings for each dimension were standardized before the distance measures were created, because all of the six analytic scores were not on the same score scale (i.e., 1–6 for the first two dimensions, 1–5 for the remaining dimensions). The Euclidean distance model was used to estimate the parameters.[4] The minimum and maximum numbers of dimensions were set at one and three, respectively, since more than three dimensions was not feasible, given the small number of the variables and sample sizes for the data. The alternating least squares scaling (ALSCAL) method (Young & Lewyckyj, 1979) implemented in SPSS was used for the optimization process. The maximum number of iterations and the convergence criterion for changes in Young's S-stress (Takane, de Leeuw, & Young, 1977) between iterations was set at 500 and 0.0001, respectively.[5]

Score profile analyses. Score profile analyses were conducted in two phases. In Phase 1, two different polytomous Rasch IRT analyses were conducted for each prompt using the computer program Facets (Linacre, 1998) to identify those examinees that had nonuniform analytic score profiles that deviated significantly from the expectation of the unidimensional Rasch IRT model. Both the partial credit model (PCM) and a PCM version of the many faceted Rasch measurement (MFRM) model (Linacre, 1989, 1998) were used to analyze the data. For the first model, only the examinees and rating dimensions were modeled as measurement facets, and the averaged scores over two raters were used as units of analyses. For the second model, all three measurement facets (examinees, raters, and rating dimensions) were modeled.[6] Misfit statistics were obtained for each of the examinees from these two separate Rasch analyses. Infit mean square values (equal to or greater than 1.4 in both analyses) were used to identify a group of examinees with nonuniform score profiles.[7]

In Phase 2, cluster analysis (CA) was done to investigate whether it is possible to obtain some meaningful or interpretable score profiles from the analytic score patterns of the misfitting examinees. When analytic scores obtained for different rating dimensions for each of the essays are highly correlated among themselves, it is very likely that most of the essays will turn out to have uniform score profiles. For this reason, the CA was intentionally done only on a smaller subsample of examinees (or essays) that were identified as having nonuniform score profiles.

CA is a statistical technique that is used to sort cases into homogeneous groups based on selected characteristics (Romesburg, 2004).The goal of CA is to minimize variability within clusters but maximize variability among clusters. The *k*-means clustering method (Anderberg, 1973) was used to determine the best number of clusters leading to the greatest separation among clusters. Several rounds of analyses were conducted for both prompts, with different numbers of clusters assumed for the examinees in each round. In each round of the analyses, a close inspection was made of the score patterns for the examinees (or essays) classified under each of the clusters. Finally, the five-examinee cluster model was selected to represent the nonuniform score profiles for both prompts.

Patterns of center means (i.e., means of analytic scores within clusters) were compared across clusters to examine the defining characteristics of each of these score clusters. In CA, the center mean for a dimension is defined to be a mean score for all the cases in each of the clusters (Romesburg, 2004). Since the six analytic scores were on slightly different score scales, the center means in each cluster were also computed based on standardized (z) scores (as well as raw ratings).

Analyses of correlations between analytic essays scores and e-rater essay features. Correlation matrices for essay scores and *e-rater* essay feature variables were obtained to examine the relationships between each of the analytic scores and the *e-rater* essay feature variables. Both Pearson zero-order and partial correlations were examined. The partial correlations were examined to see how much unique contribution each of the regular *e-rater* variables could make in predicting each of the analytic scores, independently of essay length.

Results

Generalizability Analyses of Analytic and Holistic Scores

Tables 6 and 7 show the G-study variance components for the six analytic scoring dimensions (development, organization, vocabulary, sentence variety/construction, grammar/usage, mechanics), the covariance components and universe score correlations between the rating dimensions, and the percentage of variance contributed by each variance component to the total subsection (rating dimension) variance estimated from two separate multivariate analyses ($p^\bullet \times r'^\bullet$) for Prompts 1 and 2, respectively.

Among the three variance components estimated for each scoring dimension in the two separate analyses, the largest variance component was that associated with persons [$\sigma^2(p)$], the second largest with the person-by-rating interaction plus undifferentiated error [σ^2(pr, undifferentiated)], followed by the rating main effect variances [$\sigma^2(r)$]. Overall, the person variance was by far the largest component of variance across the six rating dimensions. When the relative proportion of the person variance was compared across the six dimensions, it was largest in the vocabulary dimension (84%, 82%), but smallest in the organization dimension (69%, 69%) for both prompts. In contrast, the person-by-rating interaction plus undifferentiated error variance was largest in the organization dimension (30%, 31%), whereas it was smallest in the vocabulary dimension (16%, 18%). Please also note that the person variance becomes a universe (true) score variance in computing the score reliability coefficients, while the remaining variance components are used to define error variances.

Table 6

Estimated Variance and Covariance Components for Prompt 1

Effect	DEV		ORG		VOC		SVC		GU		MEC	
	Var/ cov	%	Var/ cov	%	Var/ cov	%	Var/ cov	%	Var/ cov	%	Var/ cov	%
Person	**1.203**	79.0	*0.99*		*0.94*		*0.90*		*0.87*		*0.78*	
(p)	1.205		**1.228**	69.3	*0.92*		*0.91*		*0.88*		*0.80*	
	1.209		1.187		**1.368**	83.7	*0.97*		*0.94*		*0.82*	
	1.197		1.224		1.366		**1.460**	79.5	*0.95*		*0.83*	
	1.045		1.066		1.202		1.255		**1.198**	71.9	*0.87*	
	0.963		1.001		1.087		1.136		1.071		**1.269**	70.2
Rating	**0.004**	0.2										
(r′)	0.008		**0.013**	0.8								
	–0.004		–0.007		**0.003**	0.2						
	0.000		–0.001		0.000		**0.000**	0.0				
	0.016		0.029		–0.015		–0.002		**0.056**	3.4		
	–0.005		–0.010		0.005		0.000		–0.019		**0.005**	0.3
Person	**0.316**	20.7										
-by-	–0.014		**0.530**	29.9								
rating	0.013		0.041		**0.264**	16.1						
(pr′)	0.033		–0.015		–0.007		**0.377**	20.5				
	0.035		0.019		0.004		0.176		**0.412**	24.7		
	–0.028		0.001		–0.036		0.018		–0.011		**0.533**	29.5
Total	**1.522**	100.0	**1.771**	100.0	**1.635**	100.0	**1.837**	100.0	**1.666**	100.0	**1.808**	100.0

Note. Boldfaced numbers are variances, numbers in the lower diagonal are covariances, and italicized numbers are correlations. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics, var = variance, cov = covariance.

Table 7

Estimated Variance and Covariance Components for Prompt 2

	DEV		ORG		VOC		SVC		GU		MEC	
Effect	Var/ Cov	%	Var/ Cov	%	Var/ Cov	%	Var/ Cov	%	Var/ Cov	%	Var/ Cov	%
Person	**1.015**	74.3	*0.95*		*0.93*		*0.92*		*0.88*		*0.85*	
(p)	0.969		**1.025**	68.8	*0.91*		*0.90*		*0.84*		*0.84*	
	1.069		1.048		**1.302**	82.3	*0.92*		*0.91*		*0.82*	
	1.104		1.088		1.252		**1.428**	79.9	*0.96*		*0.84*	
	0.977		0.941		1.150		1.270		**1.220**	75.5	*0.86*	
	0.986		0.971		1.078		1.154		1.087		**1.315**	72.8
Rating	**0.030**	2.2										
(r′)	0.011		**0.003**	0.2								
	0.009		0.003		**0.002**	0.1						
	–0.014		–0.005		–0.004		**0.005**	0.3				
	0.025		0.010		0.008		–0.012		**0.021**	1.3		
	–0.002		–0.001		–0.001		0.001		–0.002		**0.000**	0.0
Person	**0.322**	23.5										
-by-	0.033		**0.462**	31.0								
rating	–0.008		–0.003		**0.277**	17.5						
(pr′)	0.015		0.013		0.012		**0.353**	19.8				
	0.034		0.014		0.011		0.134		**0.375**	23.2		
	–0.007		0.041		0.007		–0.022		–0.005		**0.490**	27.2
Total	**1.366**	**100.0**	**1.491**	**100.0**	**1.582**	**100.0**	**1.786**	**100.0**	**1.615**	**100.0**	**1.805**	**100.0**

Note. Boldfaced numbers are variances, numbers in the lower diagonal are covariances, and italicized numbers are correlations. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics, var = variance, cov = covariance.

Universe score correlations among these six analytic scores were high. This was particularly true among the first four dimensions, which include development, organization, vocabulary, and sentence variety/construction dimensions (> 0.90). In contrast, the remaining two dimensions, especially mechanics, seemed to be somewhat distinct from the first four dimensions. The highest universe score correlation was obtained between the development and organization scores (0.99, 0.95) for both prompts, whereas the lowest correlation was observed for mechanics and development for Prompt 1 (0.78) and for mechanics and vocabulary scores for Prompt 2 (0.82).

Figure 1 displays the score reliability coefficients (i.e., dependability indices) for each of the six analytic scores and for the composite score estimated for a double-rating scheme for Prompts 1 (P1) and 2 (P2). In addition, the score reliability coefficients for the holistic score are also shown on the same figure. Similar results were obtained for both prompts. Since the rating main effect was small for each of the dimensions for the two prompts, both the dependability indices and generalizability coefficients were very close. The dependability indices (or phi coefficients) ranged from 0.81 to 0.91 across the six dimensions for the two prompts. Higher score reliability estimates were obtained for the vocabulary, development, and sentence variety/construction dimensions (0.85–0.91) than for the organization, grammar/usage, and mechanics dimensions (0.81–0.86). Overall, acceptable levels of score reliabilities (> 0.80) were achieved for both prompts across all of the six rating dimensions.

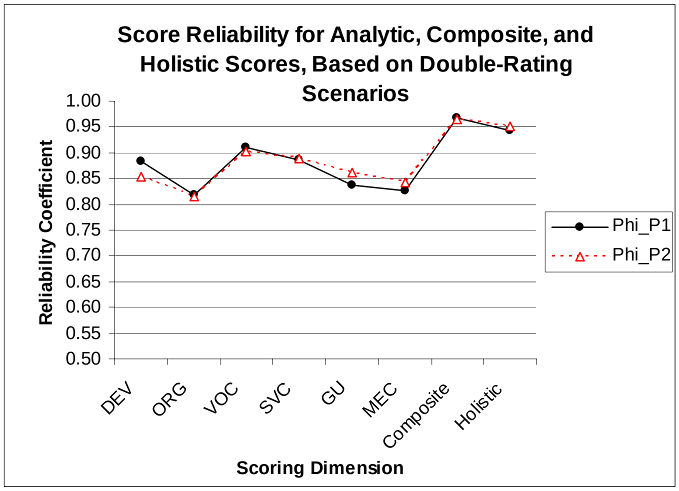

Figure 1. **Reliability of analytic and composite scores based on double ratings for each of the two TOEFL CBT prompts.**

Correlation Between Analytic and Holistic Scores

Table 8 shows the Pearson zero-order and partial correlations among the holistic score, the six analytic essay scores, and the essay length variable (TNW) for Prompts 1 and 2. In the table, the elements below the diagonal represent the zero-order correlations, while those above the diagonal represent (italicized) partial correlations.

First, in terms of the Pearson correlations, we found that the six analytic scores were correlated significantly among themselves. The highest correlation was observed for the two language-use subdimension pairs of sentence variety/construction and grammar/usage for both prompts (0.88.0.89), which was followed by the sentence variety/construction and vocabulary dimension pair (0.87, 0.83). The lowest correlations were obtained for the mechanics and development pair for the first prompt (0.66) and for the mechanics and organization pair for the second prompt (0.71).

Second, each of the six analytic scores was also correlated significantly with the holistic score for both prompts. Of the six analytic scores, the vocabulary, development, and sentence construction dimensions were most strongly correlated with the holistic scores (0.87–0.90). The organization and grammar/usage dimensions were also highly correlated with the holistic scores (0.83–0.85). The lowest correlation was observed for the mechanics dimension (0.72, 0.75).

Third, one intriguing result was that both analytic and holistic scores were significantly correlated with the essay length variable measured by the total number of words in an essay (TNW). Above all, the holistic score was more highly correlated with essay length (0.89, 0.90) than any of the six analytic scores (0.60–0.88). When only the six analytic scores were compared, the first four dimensions (development, organization, vocabulary, and sentence variety/construction) were more sensitive to essay length than were the last two dimensions (grammar/usage and mechanics). More specifically, the development score was most strongly correlated with the essay length variable (0.88), while the mechanics score was most weakly correlated (0.60, 0.67).

Table 8 also shows partial correlations among the averaged holistic and analytic scores for Prompts 1 and 2, after the linear and quadratic impact of essay length was removed. A close inspection of partial correlations (elements above the diagonal in the tables) revealed that all of the six analytic scores were correlated among themselves at a significant, but much lower, level after the impact of essay length was controlled. The highest partial correlation was obtained between the two language-use subdimension scores of sentence variety/construction and grammar/usage (0.67, 0.69) for both prompts. This can be considered high, given that the partial correlations for the remaining pairs ranged from 0.22 to 0.50 for the first prompt and from 0.15 to 0.45 for the second prompt.

Table 8

Pearson and Partial Correlations Between Holistic and Analytic Scores

			Analytic					
		Holistic	DEV	ORG	VOC	SVC	GU	MEC
Prompt 1								
Holistic		**1.00**	*0.35*	*0.35*	*0.50*	*0.49*	*0.47*	*0.40*
Analytic	DEV	0.88	**1.00**	*0.39*	*0.35*	*0.29*	*0.29*	*0.24*
	ORG	0.85	0.84	**1.00**	*0.25*	*0.22*	*0.25*	*0.24*
	VOC	0.90	0.85	0.81	**1.00**	*0.50*	*0.49*	*0.34*
	SVC	0.87	0.81	0.78	0.87	**1.00**	*0.67*	*0.39*
	GU	0.83	0.77	0.74	0.83	0.88	**1.00**	*0.45*
	MEC	0.72	0.66	0.66	0.70	0.72	0.73	**1.00**
TNW		0.89	0.88	0.80	0.84	0.79	0.75	0.60
Prompt 2								
Holistic		**1.00**	*0.35*	*0.24*	*0.44*	*0.48*	*0.55*	*0.28*
Analytic	DEV	0.88	**1.00**	*0.25*	*0.25*	*0.26*	*0.29*	*0.27*
	ORG	0.83	0.81	**1.00**	*0.19*	*0.18*	*0.15*$^{+}$	*0.20*
	VOC	0.88	0.82	0.78	**1.00**	*0.38*	*0.45*	*0.26*
	SVC	0.87	0.81	0.77	0.83	**1.00**	*0.69*	*0.26*
	GU	0.85	0.77	0.71	0.81	0.89	**1.00**	*0.37*
	MEC	0.75	0.73	0.71	0.72	0.72	0.73	**1.00**
TNW		0.90	0.88	0.80	0.82	0.80	0.74	0.67

Note. N = 465. Boldface numbers indicate the diagonal. Elements below the diagonal are original Pearson correlations. Elements above the diagonal (italicized) are partial correlations. All of the correlation coefficients were statistically significant at the 0.05 level (two-tailed) and all except one ($^{+}$) also are significant at the 0.01 level. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics, TNW = total number of words.

Moreover, we also found that all of the six analytic scores were still correlated with the holistic scores at a weak but statistically significant level (0.24–0.55), after the impact of essay length was removed. It was the vocabulary, sentence variety/construction, and grammar/usage scores that were more highly correlated with the holistic scores (0.44–0.55). The correlations between the holistic score and three analytic scores of development, organization, and mechanics were lower (0.24–0.40). This suggests that the three language-related dimensions (vocabulary, sentence variety/construction, grammar/usage) have greater, essay-length independent, explanatory power for the holistic scores than the development and organization dimensions (development, organization).

Multidimensional Scaling Analyses of Holistic and Analytic Scores

To examine further the empirical relationships among the holistic and analytic scores through a graphical representation of these scores, multidimensional scaling analyses were conducted separately for each of the two prompts. Figure 2 shows the plots of the holistic and analytic scores in a two-dimensional space obtained from the multidimensional scaling analysis (see also Appendix C for results for single- and three-dimensional solutions). Similar results were obtained for both prompts, as shown in Figure 2. First, the mechanics score (MEC) seemed to be somewhat distinct from the remaining six scores for both prompts. The first dimension represented by the X-axis (abscissa) seemed to be playing an important role in separating the mechanics score from the remaining six scores. The mechanics score is located horizontally on the far left (negative) side of Dimension 1, whereas the remaining six scores were scattered around the midpoint on the first dimension, leaning more toward the positive side.

Second, the remaining six scores were differentiated vertically on the second dimension represented by the Y-axis (ordinate). For both prompts, holistic scores (HOL) were located near the midpoint on the second dimension. Located above the holistic score (on the positive side) are the three language-related dimensions of vocabulary (VOC), sentence variety/construction (SVC), and grammar/usage (GU), whereas the development (DEV) and organization (ORG) dimensions were located below the holistic score (on the negative side). The two variables that were closest to the holistic score in distance on Dimension 2 were the vocabulary and development scores on the positive and negative sides, respectively, for Prompt 1. For Prompt 2, however, the vocabulary and the sentence variety/construction scores on the positive side were located closest to the holistic score.

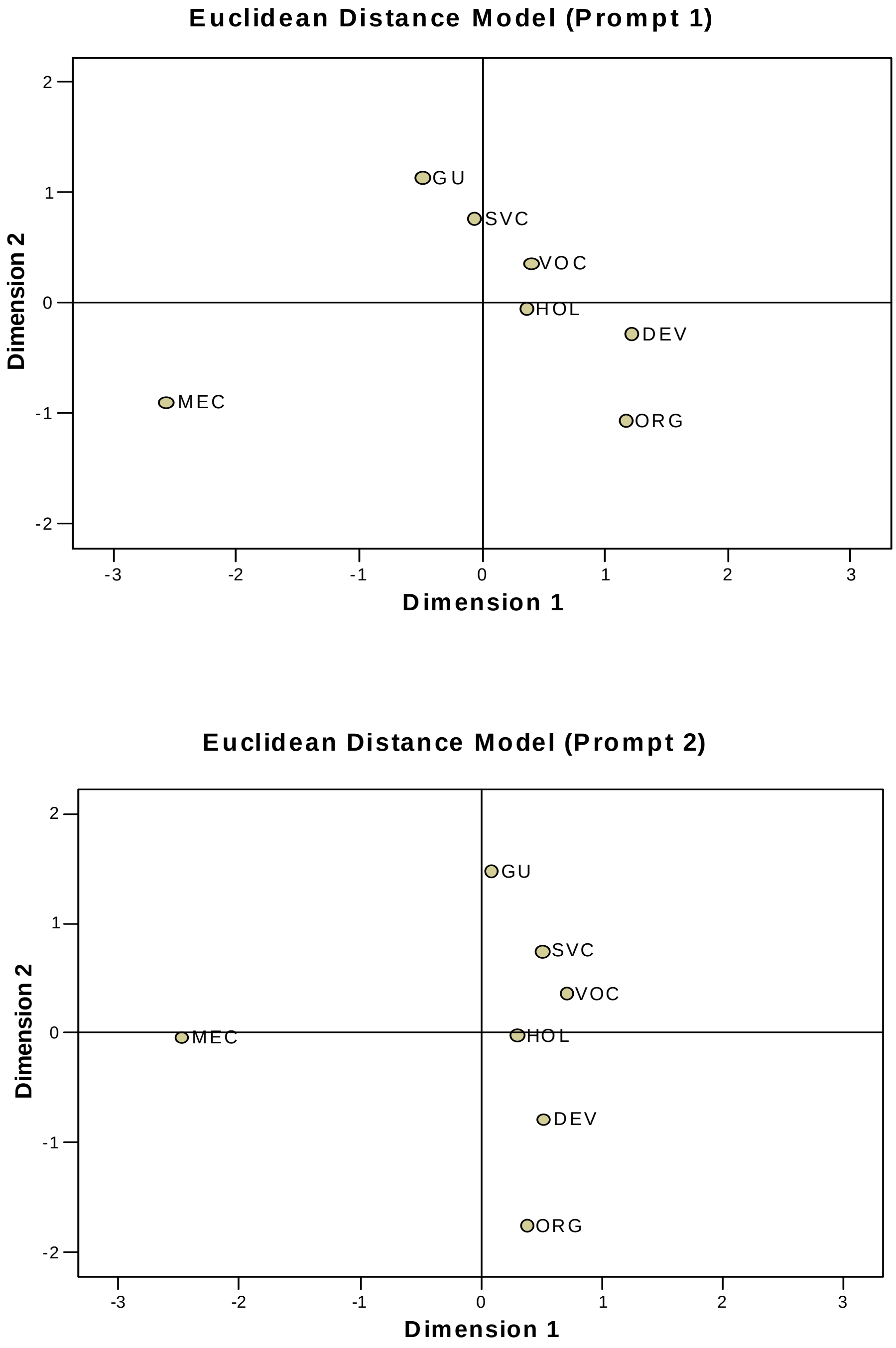

***Figure 2.* Representation of holistic and analytic scores in the two-dimensional space based on multidimensional scaling analysis for two writing prompts.**

Score Profile Analyses

For score profile analyses, we first selected the examinees who had significantly large infit mean square values (> = 1.4) in both PCM and MFRM analyses. For Prompt 1, 61 of 465 examinees (about 13%) were identified as having unusually large examinee misfit statistics, while a total of 51 examinees (about 11%) were identified as misfitting for Prompt 2. For each of the two prompts, the misfitting examinees were classified into five different clusters based on CA. The purpose of the CA was to investigate whether it is possible to obtain some interpretable score profiles (or examinee clusters) from the analytic score patterns of the misfitting examinees.

Tables 9 and 10 show the number of examinees classified into each of the five clusters and the center means for both raw ratings and standardized scores for each of the six analytic rating dimensions on each cluster for the two prompts (see Tables D1 and D2 for examinee score profiles for these clusters). Figures 3 and 4 also show the patterns of the standardized mean scores for each cluster graphically. As shown in Tables 9 and 10, it seems that there are some noticeable similarities among clusters identified for Prompts 1 and 2 in terms of defining characteristics of the clusters, although there were some subtle differences as well.

As expected, the mechanics score seemed to play an important role in defining some of the clusters identified for each of the two prompts (e.g., Cluster 5 for both prompts, Cluster 1 for Prompt 1, and Cluster 3 for Prompt 2).First, Cluster 5 for both prompts was similar in that these two clusters were characterized by high development/organization, high vocabulary/sentence variety/grammatical accuracy, and low mechanical accuracy scores overall. In these two clusters, examinees’ standardized analytic scores on the first five rating dimensions are comparatively high (0.7–1.4), whereas the mechanics score is significantly lower (–0.4—0.1). Such a score pattern seems to happen mostly in relatively long, well-developed essays. The averaged TNW for essays in these two clusters were 371 and 345, respectively, for the two prompts. Another distinctive nature of the essays belonging to Cluster 5 for both prompts is that they tended to receive high holistic scores, since mechanical accuracy plays a minor role in the holistic judgment of essay quality for most of the essays. The average holistic scores for essays in Cluster 5 were 5.2 and 5.1 for Prompts 1 and 2, respectively. (The average standardized holistic score for the cluster was 1.1 for both prompts.)

Table 9

Means of Six Analytic Scores in Each of the Five Score Clusters for Prompt 1

Rating dimension	Cluster number									
	1 (*n* = 18)		2 (*n* = 10)		3 (*n* = 10)		4 (*n* = 9)		5 (*n* = 14)	
	Raw	SD	Raw	SD	Raw	SD	Raw	SD	Raw	SD
DEV	2.9	–0.6	5.0	1.2	2.1	–1.3	3.9	0.3	5.3	1.4
ORG	3.2	–0.5	5.6	1.5	1.7	–1.7	4.6	0.6	5.1	1.0
VOC	2.3	–0.6	3.8	0.5	2.0	–0.9	3.8	0.6	4.4	1.1
SVC	2.0	–0.8	3.2	0.1	1.9	–0.9	4.2	0.9	4.5	1.2
GU	2.1	–0.6	2.7	–0.1	1.7	–0.9	4.2	1.2	4.1	1.1
MEC	3.9	0.7	3.2	0.1	2.0	–0.9	4.8	1.3	2.6	–0.4
HOL	2.8	–0.6	4.7	0.8	1.9	–1.3	4.4	0.6	5.2	1.1
TNW	131	–0.7	318	1.1	98	–1.1	224	0.2	371	1.6

Note. Based only on 61 misfitting examinees. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics, HOL = holistic, TNW = total number of words.

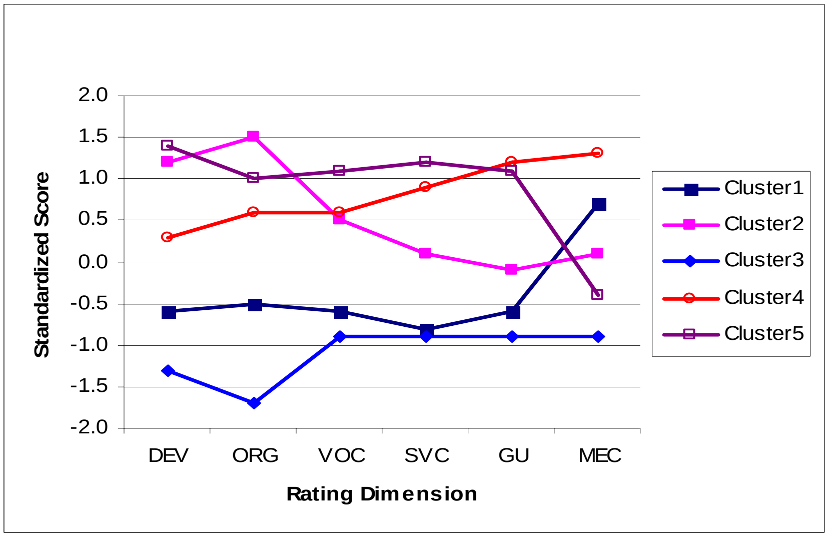

***Figure 3.* Patterns of means of six analytic scores in each of the five score clusters for Prompt 1.**

Table 10

Means of Six Analytic Scores in Each of the Five Score Clusters for Prompt 2

	Cluster number									
Rating dimension	1 (*n* = 8)		2 (*n* = 8)		3 (*n* = 15)		4 (*n* = 10)		5 (*n* = 10)	
	Raw	SD	Raw	SD	Raw	SD	Raw	SD	Raw	SD
DEV	3.3	–0.4	2.4	–1.2	3.2	–0.4	4.8	1.1	4.8	1.0
ORG	3.9	0.1	2.3	–1.4	3.4	–0.4	5.5	1.5	4.7	0.7
VOC	3.6	0.4	1.4	–1.3	2.4	–0.5	4.2	0.9	4.2	1.0
SVC	4.1	0.8	1.8	–1.0	2.2	–0.7	3.4	0.3	4.6	1.2
GU	2.9	0.1	1.6	–1.0	2.2	–0.5	2.5	–0.2	4.7	1.6
MEC	2.4	–0.7	2.1	–1.0	4.3	0.8	4.2	0.7	3.4	0.1
HOL	3.9	0.2	1.8	–1.3	3.1	–0.4	4.8	0.8	5.1	1.1
TNW	223	0.1	84	–1.2	153	–0.6	357	1.4	345	1.2

Note. Based only on 51 misfitting examinees. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics, HOL = holistic, TNW = total number of words.

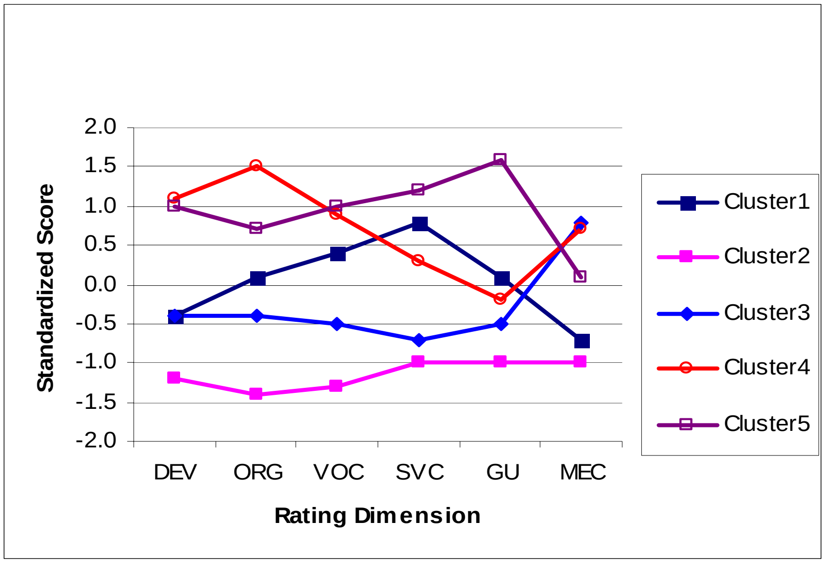

Figure 4. **Patterns of means of six analytic scores in each of the five score clusters for Prompt 2.**

Second, Cluster 1 for Prompt 1 and Cluster 3 for Prompt 2 were also similar in that these two clusters were characterized by medium development/organization, medium vocabulary/sentence variety/grammatical accuracy, and high mechanical accuracy scores. It should be noted that these two clusters represent a score pattern almost opposite to that of Cluster 5 mentioned above. In these two clusters, the standardized mechanics score was the highest one (0.7, 0.8) among the six analytic scores, while the rest of the scores were in the low score to midscore point range (–0.8 — –0.4). Such a score pattern seemed to happen in relatively short or medium-length essays (The averaged TNW for essays in these two clusters were 131 and 153, respectively, for the two prompts.) As previously mentioned, mechanics plays only a limited role in the holistic rating of essays, and these examinees tended to receive holistic scores that were more similar to development scores and other nonmechanics scores. The average holistic scores for essays in these clusters were 2.8 and 3.1, respectively, for the two prompts. (The average standardized holistic scores for these clusters were –0.6 and –0.4, respectively.)

In addition to the above mentioned three clusters defined by the mechanics versus nonmechanics distinction, there were some other noteworthy clusters that can be better defined by somewhat different characteristics. These three clusters include Cluster 2 for Prompt 1, Cluster 4 for Prompt 2, and Cluster 4 for Prompt 1. All of these clusters seem to represent an interesting score profile in terms of the content versus language distinction often made in the ESL writing literature (Cumming et al., 2002; Santos, 1988). The first cluster (Cluster 2 for Prompt 1), for instance, can best be characterized by high development/organization scores and medium vocabulary/sentence variety/ grammar/mechanics scores. The second cluster (Cluster 4 for Prompt 2) showed a similar score pattern, although relatively higher vocabulary and mechanical accuracy scores were observed for the second cluster. Such score patterns also seemed to happen mostly in relatively long essays. The average TNW for the essays classified under these two clusters were 318 and 357, respectively, for the two prompts. The average holistic scores for the essays in these two clusters were 4.7 and 4.8, respectively, for the two prompts. Another noteworthy characteristic is that the mean standardized organization score (1.5, 1.5) also tended to be somewhat higher than the mean standardized development score (1.2, 1.1) in the two clusters.

In contrast, the third cluster (Cluster 4 for Prompt 1) was represented by medium writing development/organization, high sentence variety/grammatical and mechanical accuracy. This score pattern represents almost an opposite score pattern to Cluster 2 for Prompt 1 mentioned above, except for the fact that the vocabulary score was similar in both clusters. Such score patterns also seemed to happen mostly in medium (close to average) length essays. The average TNW for the essays classified under these two clusters was 224. The average holistic scores for the essays in these two clusters was 4.4. (The averaged standardized holistic score was 0.6.) In a nutshell, the defining characteristics of these three clusters seem to provide some partial support for the distinction between the content/rhetoric and language control aspects of ESL writing often made by ESL writing researchers (Cumming et al., 2002; Santos, 1988).

Correlation Analyses of Analytic Scores and E-rater Essay Feature Variable Scores

Table 11 displays averaged Pearson (or zero-order) correlations and Table 12 displays averaged partial correlations between the six analytic scores and 12 *e-rater* essay feature variables across the two prompts. Since the results for Prompts 1 and 2 are similar, only averaged correlations across the two prompts are reported here.

Pearson correlations. Table 11 shows the averaged Pearson correlations between the seven essay scores and the 12 *e-rater* essay feature scores across the two prompts. First, the essay length (TNW) turned out to be the automated essay feature variable that had the strongest correlation with all of the six analytic scores (0.64–0.88) and the holistic score (0.90). Among them, the holistic and development scores were most sensitive to the essay length variable in particular. Nevertheless, it should be noted that the strength of the relationship between the essay length and the grammar (0.74) and mechanics scores (0.64) seemed somewhat weaker than those between the essay length and the rest of the analytic and holistic scores.

Second, of the two discourse and organization features used in *e-rater,* the discourse unit score seemed to have consistently moderate correlations with all of the seven essay scores (0.57–0.70). This essay feature was correlated more strongly with the first three human supplied scores (holistic, DEV, ORG) than with the remaining four analytic scores (VOC, SVC, GU, MEC). This is a somewhat expected result, given that this feature is intended to tap into the organizational aspect of essay quality. However, the average length of discourse units variable had much lower correlations with the seven essay scores (0.16–0.27).

Third, among the three lexical complexity variables, the vocabulary level feature based on word-frequency levels had consistently moderate correlations with the seven scores for both prompts (0.40–0.60). In relation to this feature, one encouraging finding was that this vocabulary level variable was most highly correlated with the vocabulary (VOC) and holistic scores (HOL) assigned by human raters (0.60). In contrast, the type-token ratio and word length variables had much smaller correlations with any of the seven essay scores.

Table 11

Averaged Pearson Correlations Between Essay Scores and E-rater Essay Feature Scores Across the Two Prompts

E-rater variables	Holistic	Analytic					
		DEV	ORG	VOC	SVC	GU	MEC
Discourse unit score	0.69	0.68	0.70	0.63	0.63	0.58	0.57
Length of discourse unit	0.26	0.24	0.20	0.27	0.25	0.22	0.16
Type/token ratio	0.36	0.39	0.45	0.27	0.31	0.29	0.37
Word length	0.13	0.11	0.14	0.14	0.13	0.12	0.10
Vocabulary level	0.60	0.58	0.52	0.60	0.56	0.52	0.40
Word-vector score	0.53	0.45	0.43	0.54	0.52	0.49	0.35
Word-vector correlation	0.77	0.71	0.67	0.71	0.70	0.68	0.57
Grammatical accuracy ratio	0.39	0.35	0.34	0.38	0.38	0.37	0.37
Usage accuracy ratio	0.11	0.09	0.12	0.11	0.10	0.12	0.18
Stylistic accuracy ratio	0.62	0.58	0.52	0.65	0.60	0.55	0.46
Mechanical accuracy ratio	0.50	0.44	0.46	0.45	0.46	0.46	0.59
Essay length	0.90	0.88	0.80	0.83	0.80	0.74	0.64

Note. N = 930; All correlations were significant at the .01 level (two-tailed). DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics, HOL = holistic, TNW = total number of words.

Fourth, of the two prompt-specific vocabulary usage variables, the word-vector (cosine) correlation variable had high correlations with the seven essay scores (0.57–0.77). This content-word vector correlation variable was most strongly correlated with the holistic score and most weakly with the mechanics score. The other content-word vector variable (i.e., the word vector score) also had significant correlations with all of the seven essay scores (0.35–0.54), although the magnitude of correlation was somewhat smaller for this word vector score variable overall.

Another thing to note is that the word vector score was most strongly correlated with the vocabulary (0.54) and holistic scores (0.53).

Table 12

Partial Correlations Between Essay Scores and E-rater Essay Feature Scores for the Two Prompts

E-rater variables	Holistic	Analytic					
		DEV	ORG	VOC	SVC	GU	MEC
Discourse unit score	0.11	0.14	0.23	0.00[a]	0.06[a]	0.05[a]	0.11
Length of discourse unit	–0.14	–0.15	–0.18	–0.04[a]	–0.07[a]	–0.08[a]	–0.11
Type/token ratio	–0.24	–0.09[a]	0.05[a]	–0.37	–0.23	–0.18	0.01[a]
Word length	0.38	0.29	0.30	0.32	0.26	0.21	0.15
Vocabulary level	0.07[a]	0.05[a]	–0.03[a]	0.15	0.06[a]	0.07[a]	–0.07[a]
Word-vector score	0.30	0.08	0.09	0.31	0.28	0.23	0.06[a]
Word-vector correlation	0.20	0.02[a]	0.04[a]	0.10	0.15	0.19	0.09[a]
Grammatical accuracy ratio	0.20	0.11	0.10	0.17	0.18	0.17	0.17
Usage accuracy ratio	0.18	0.11	0.16	0.14	0.13	0.15	0.22
Stylistic accuracy ratio	0.14	0.06[a]	–0.02[a]	0.28	0.17	0.12	0.04[a]
Mechanical accuracy ratio	0.30	0.15	0.18	0.16	0.19	0.22	0.43

Note. Essay length was not applicable. DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, and MEC = mechanics.

[a]Correlations were not significant at the .05 level.

Fifth, among the four linguistic accuracy variables, the stylistic accuracy ratio variable was consistently most highly correlated with the seven essay scores (0.46–0.65), with it being most highly correlated to the vocabulary score. Since one major type of errors that contribute largely to the stylistic accuracy ratio in *e-rater* is excessively repeated words in the essay, the highest correlation between the stylistic accuracy ratio and the vocabulary score was somewhat expected. The grammatical accuracy ratio was also correlated with the seven essay scores at a significant level (0.35–0.39). In contrast, the usage accuracy ratio had lower correlations with the seven essay scores (0.09–0.18). This was also an expected result, given that there was only a very small variation in the usage accuracy ratio across the essays (see Table 5).

In relation to the linguistic accuracy variables, the most intriguing finding was that a close link was confirmed between the mechanical accuracy ratio computed by *e-rater* and the mechanical accuracy rating (MEC) given by human raters. The mechanical accuracy ratio was correlated with all of the seven essay scores (0.44–0.59). More importantly, the mechanical accuracy ratio was most strongly correlated with the mechanics score assigned by human raters.

Partial correlations. Table 12 shows the partial correlations between the seven essay scores and the 12 *e-rater* essay feature scores across the two prompts. The partial correlations are reported here to show how much unique contribution each of the 11 *e-rater* variables (except essay length) can make in predicting each of the analytic scores, independently of essay length. As expected, the magnitude of partial correlations between the human-assigned essay scores and the 11 *e-rater* essay feature variables were much smaller than that of the zero-order Pearson correlations between these variables.

First, of the two variables related to discourse and organization that are used in *e-rater*, the discourse unit score had positive partial correlations with all of the seven essay scores, including both the holistic and analytic scores (0.00–0.23). The discourse unit score again turned out to have the strongest, positive correlation with the organization score assigned by human raters. As mentioned previously, the discourse unit score can be related conceptually to the organizational aspect of essay quality. In contrast, the average length of discourse units had negative, partial correlations with all of the seven essay scores (–0.18 — –0.04). More intriguingly, it had the strongest negative correlation with the organization score. Second, when the three lexical complexity variables were compared, the average word length turned out to be the essay feature that had the highest positive partial correlation with all of the seven essay scores (0.15–0.38). Even when the partial correlations were compared across the 11*e-rater* variables, the average word length variable had the highest positive correlation with the first five essay scores assigned by humans (HOL, DEV, ORG, VOC, and SVC). In contrast, the vocabulary level variable, although it had the highest zero-order correlation with the seven essays scores in the previous analysis, turned out to have consistently lower partial correlations with the seven essay scores than the average word length (–0.07–0.15). Nevertheless, the vocabulary level feature variable was again mostly highly correlated with the vocabulary score assigned by humans (0.15). A rather unexpected finding was that the type-token ratio had negative partial correlations with most of these essays scores assigned by human raters, particularly with the vocabulary score assigned by human raters (–0.37).

Third, both of the prompt-specific vocabulary usage variables had small correlations with the seven essay scores. The word vector score (0.06–0.31) had somewhat higher correlations with all of the six scores (except the mechanics score) than the word vector correlation (0.02 0.20). We should be reminded, however, that an opposite pattern was observed for the zero-order Pearson correlations in the previous analysis. One noteworthy finding was that the word vector score once again had the highest correlation with the vocabulary rating (0.31) given by human raters, even when the partial correlation was examined.

Fourth, among the four linguistic accuracy variables, the mechanical accuracy ratio was positively correlated with all of the seven essay scores (0.15 0.43), even after the impact of essay length on the correlation was removed. Particularly, this mechanical accuracy ratio had the strongest, positive correlation with the mechanics rating (MEC) by human raters (0.43). This suggests that what is tapped by the mechanical accuracy ratio may be closely related to the human raters' judgment of essay quality in terms of mechanical accuracy in the examinee's essays. In addition, the stylistic accuracy ratio turned out to be most strongly correlated to the vocabulary scores (0.28) again, whereas the grammatical accuracy ratio had positive correlations with all of the seven scores.

Summary and Discussion

The main purposes of the study were to develop and evaluate analytic scoring rubrics for TOEFL CBT writing prompts, to examine the relationships among holistic and analytic essay scores, and to investigate relationships between analytic scores and *e-rater* essay feature scores. One important additional objective was to investigate ways to generate meaningful profile scores based on these analytic scores and link them to *e-rater* essay features. High score reliability was achieved for all of these six analytic dimensions. It was found that (a) all of the six analytic scores were not only correlated among themselves but also correlated with the holistic scores, (b) high correlations obtained among holistic and analytic scores were largely attributable to the impact of essay length on both analytic and holistic scoring, (c) there may be some potential for profile scoring based on analytic scores, and (d) some strong associations were confirmed between several *e-rater* variables and analytic ratings. These findings are discussed next in more detail.

Relationships Among Analytic Scores and Holistic Scores

Close examinations of zero-order Pearson product moment correlations and MDS results revealed that, although all of the seven analytic and holistic dimensions were correlated among themselves at moderate to high levels, they seemed to be measuring related but somewhat distinct aspects of essay quality. First, all of the six analytic scores were significantly correlated among themselves, with the strength of the relationship varying across pairs of dimensions. Looking across the prompts that were analyzed, the highest correlations were observed between the two subdimensions of language use (i.e., sentence variety/construction and grammar/usage accuracy). The next highest correlations were observed between the sentence variety/construction and vocabulary dimensions. The lowest correlations were obtained for the mechanics and development pair for the first prompt and for the mechanics and organization pair for the second prompt.

Second, we also found that each of the six analytic scores was correlated with holistic scores, with the strength of relationships again varying across the dimensions. The development, vocabulary, and sentence variety/construction scores were most strongly correlated with the holistic scores. The organization and grammar/usage scores were also highly correlated with holistic scores. The lowest correlations were observed for mechanics. The mechanics dimension seems to be most distinct from the rest of the dimensions for which analytic scores were obtained, whereas the vocabulary and development scores seem to be most closely related to the holistic score. These results are consistent with previous research findings on the relationships between analytic and holistic ratings assigned to ESL learner’s essays (Bacha, 2001) and those between lexical diversity and holistic scores (Engber, 1995; Laufer & Nation, 1995). Third, MDS analyses were very helpful in further investigating the relationships among the analytic and holistic rating dimensions in more depth. An inspection of two-dimensional MDS plots (Figure 2) showed that the mechanics dimension (MEC) was most distinct from the remaining six rating dimensions. It was also found that the holistic rating dimension was very helpful in grouping the remaining five analytic dimensions (except mechanics) into two theoretically distinct clusters of dimensions in the MDS plots. The holistic score was consistently located midway between the two content/rhetoric dimensions (i.e., development, organization) and the three language-related dimensions (i.e., grammar/usage, sentence variety/construction, vocabulary) on the vertical axis in the two MDS plots. The relative distance, and order, of each of these five analytic dimensions with respect to the holistic dimension was relatively consistent across the two prompts. This suggests that the holistic score does reflect both the content-related and language-related qualities of the essays, as defined in the TOEFL CBT scoring rubric, and that the two content/rhetoric dimensions are separable to some extent from the language-related dimensions, as pointed out by some ESL writing researchers (Santos, 1988; Cumming et al., 2002).

The main implication of these findings is that, by virtue of its distinctiveness from other scores, a separate mechanics score seems clearly justified in any effort to provide a set of analytic scores. The creation of superordinate rating dimensions of content and language for profile scoring is also an additional area of research deserving further investigation. From these results, however, the justification for other analytic scores seems to be somewhat more equivocal.

Role of Essay Length in Holistic and Analytic Holistic Scores

A strong empirical relationship, not only between the essay length and holistic score but also between essay length and each of the six analytic scores used, was confirmed in this study. This was not completely unexpected, given previous research findings on the strong relationships between essay length and holistic scores (Carson, Bridgeman, Camp, & Waanders, 1985; Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Grant & Ginther, 2000; Jarvis, 2002; Jarvis et al., 2003; Reid, 1986) and between lexical diversity measures and holistic scores (Engber, 1995; Laufer & Nation, 1995). Interestingly, the holistic and development scores were found to be most highly correlated to essay length, while the mechanics score was least correlated to essay length. When only the six analytic dimensions were compared, a general tendency was that the first four, development/structure/variety-related dimensions (development, organization, vocabulary, and sentence variety/construction), tended to be strongly correlated with the essay length. However, the two accuracy-related rating dimensions (grammar/usage, mechanics) were less related to essay length, as expected.

In this regard, one important thing to mention here is the essay-length dependent nature of analytic scoring done in this study, particularly for short essays. In this study, raters were instructed, that if they saw extremely short essays of less than about 90 words (or 8 full typed lines of text), they were to assign the lowest possible score to such essays on any of the six analytic dimensions. Such a scoring rule had to be implemented in analytic scoring, because often not enough evidence or substance was found in such essays to judge the quality of the texts. Thus, it is possible that such a rating rule could have potentially contributed to the high correlations observed among the analytic scores and between the essay length and analytic scores. For this reason, the Pearson correlations obtained for the analytic scores should be interpreted with this factor in mind.

To better understand the empirical, essay-length independent relationships between the holistic and analytic essay scores, partial correlations were also computed in this study after removing the effect of essay length from the original correlations. The total number of words (TNW) and the squared values of TNW were used as covariates in computing the partial correlations to control for the linear and quadratic effects of essay length. The obtained partial correlations were significantly lower than the original correlations, but all of the six analytic scores remained correlated among themselves at a significant level, even after the impact of essay length was controlled. For both prompts, the highest partial correlation was again between the sentence variety/construction and grammar/usage scores. All six analytic scores were also correlated with the holistic scores at a significant level after the impact of essay length was removed.

One interesting pattern emerged from the partial correlations among the holistic and analytic scores, however. When zero-order correlations were compared, the vocabulary and development scores were most strongly correlated with the holistic scores. When the partial correlations were examined, however, the vocabulary score was still more highly correlated with the holistic score than were other analytic scores, but the development score was no longer correlated most strongly with the holistic score. Instead, the three dimensions related to knowledge of language components (vocabulary, sentence variety/construction, and grammar/usage) were now more highly correlated with the holistic scores than the development and organization scores. This suggests that the three language-related dimensions of vocabulary, sentence construction/variety, and grammar/usage have greater essay-length independent, explanatory power for the holistic scores than the development and organization dimensions.

The main implication of these findings is that if essay length could be controlled or constrained analytical ratings might have greater distinctiveness and therefore greater utility. This is an issue that could be researched.

Possibility of Profile Scoring

Results of this study showed that the analytic scores did not contribute as much as desired to further discrimination of examinees beyond what the holistic score can do already. As discussed earlier, all of the six analytic scores were correlated not only among themselves but also with the holistic score at a significant level in this study. The Rasch IRT analyses also revealed that most of the examinees turned out to have uniform score profiles across the six rating dimensions. Based on such findings, one could easily argue that it would be very difficult to justify the usefulness of reporting analytic scores from a statistical and psychometric point of view.

Nevertheless, it does not mean that there is no potential for profile scoring based on these analytic scores. The mechanics score was found to be less highly correlated to the holistic scores and other analytic scores. Results of MDS analyses also showed that the six analytic dimensions are tapping into related but somewhat distinct constructs. Consistent with these observations, more than 10% of examinees for each of the prompts were identified by Rasch analyses as having a nonuniform score profile. It was also found that high correlations among holistic and analytic scores were largely attributable to the impact of essay length on both analytic and holistic scoring. For this reason, the language-related dimensions (including vocabulary, sentence variety, and grammar/usage accuracy) seemed to make significant essay-length independent contributions to the holistic judgment of essay quality.

A close inspection of score profiles for essays classified into different clusters also showed that the unique score profiles for each examinee cluster might prove useful in generating feedback information about students’ essays. In one cluster for both prompts, the mechanics versus nonmechanics distinction played an important role in defining the score profile. In other clusters, the content versus language control distinction used in ESL writing was useful in identifying the defining characteristics of these clusters. In this respect, one interesting possibility is to regroup the existing six analytic rating dimensions into a somewhat smaller number of meaningful superordinate categories (e.g., content, language, and mechanics). Further investigation in this line of research will prove helpful in designing profile scoring systems based on the analytic scores.

Relationships Between Analytic Scores and E-rater Essay Features

A total of 12 essay feature variables used in *e-rater* 2.0 were analyzed in this study, which included two discourse and organization variables, three lexical complexity variables, two prompt-specific vocabulary usage variables, four linguistic and mechanical accuracy variables, and one essay length variable (TNW). Among these 12 variables, essay length (measured by the total number of words) turned out to be the strongest predictor of each of the six analytic scores as well as the holistic scores. Since the role of essay length for the holistic and analytic scores was already discussed in the previous section, the remainder of the discussion is focused on the remaining eleven variables.

First, of the two development/organization features, the discourse unit score seemed to be working as desired in tapping the surface level of organizational quality of essays. The discourse unit score had moderate-to-high correlations with all six analytic scores, with it being most strongly correlated to the organization score assigned by human raters. The discourse unit score also had the strongest positive partial correlation with the organization score. Note that the discourse unit score is defined as the difference between the actual and optimal number of discourse units in the essay, which is related to the organizational aspect of essay quality. However, the average length of discourse units in the essay turned out to have relatively lower correlations, and even negative partial correlations, with all seven essay scores. Further research seems necessary to develop a more sophisticated essay feature variable that can capture the depth of development in the essay.

Second, among the three lexical complexity variables, both the vocabulary level and average word length variables seemed to be able to capture what human raters value in terms of the lexical variety/sophistication aspect of essay quality. The vocabulary level variable consistently had moderate correlations with the six analytic scores. More importantly, this variable had the highest positive zero-order (Pearson) correlation and partial correlations with the vocabulary score. In the case of the average word length variable, this variable initially had lower Pearson correlations with the six analytic scores, but a very different pattern emerged when the partial correlations were examined. Among the three lexical variables, the word length variable had the highest partial correlation with all of the seven essays scores. In contrast, the type/token ratio had negative partial correlations with most of these essays scores assigned by human raters. Particularly, it had the highest negative partial correlation with the vocabulary score assigned by human raters (–0.37). In relation to lexical diversity measures, one interesting research area deserving further investigation is using more sophisticated type-token ratio measures that are not dependent on text length (Jarvis, 2002).

Third, both word vector variables (i.e., score, correlation) were found to be good predictors of the analytic and holistic essay scores, although the word vector correlation seemed to be more sensitive to essay length. The word vector correlation had moderate correlations with all of the seven essay scores, and it was most strongly correlated with the holistic score. The word vector score also had statistically significant correlations with the seven analytic scores, but the magnitude of correlation was somewhat smaller for the word vector score than for the word vector correlation. However, it should be mentioned that, when the partial correlations were compared, almost an opposite pattern was observed. The word vector score had higher partial correlations with the six essay scores (except the mechanics) than the word vector correlation. Besides, the word vector score was most strongly correlated with the human-assigned vocabulary score, regardless of whether the zero-order and partial correlations were used.

Lastly, the most clear-cut finding from the *e-rater* variable analysis was that it was possible to establish a link between the mechanical accuracy ratio computed by the automated scoring engine and the mechanics score (MEC) assigned by human raters. This pattern was observed both for the zero-order and partial correlations. The mechanical accuracy ratio was correlated with all of the seven essay scores, but it was most strongly correlated with the mechanics score for both prompts. The mechanical accuracy ratio also had the strongest, positive partial correlations with the mechanics score given by human raters across both prompts. This suggests that, in terms of mechanical accuracy, the mechanical accuracy ratio may reflect the same qualities that human raters attend to in examinee’s essays.

Nevertheless, we were not able to confirm a similar link between the grammatical accuracy ratio variable and the grammar/usage score assigned by human raters or between the usage accuracy ratio and the grammar/usage score. Interestingly, the stylistic accuracy ratio was most strongly correlated with the vocabulary score, whether the Pearson or partial correlations were used. In a sense, the highest correlation between the stylistic accuracy ratio and the vocabulary score is somewhat expected, given that one major type of error that contributes to the stylistic accuracy ratio is excessively repeated words across sentences and passage in the essay.

Overall, we saw reasonably strong associations between several *e-rater* variables and analytic ratings in some areas of essay quality, such as organization, vocabulary, and mechanics. This means that, for these variables, both *e-rater* and human raters are focusing on similar or related aspects of examinees’ essays. This provides some evidence supporting the validity of not only the automated text features but also the automated holistic scores computed based on these features in *e-rater*. In addition, it seems justifiable, to some extent, to use some of the existing *e-rater* variables to compute automated trait scores representing these different aspects of essay quality in addition to the automated holistic scores. However, we also noticed some conceptual mismatch between the six analytic scores and the 12 *e-rater* essay feature variables. For instance, there is clearly no *e-rater* variable that captures directly the sentence variety/construction aspect of essay quality. Further research is necessary to create the *e-rater* essay feature variables to capture a full range of essay quality features (e.g., depth of development, coherence, and appropriateness of lexical choice) valued in ESL writing.

Conclusions and Avenues for Future Research

Conclusions

In this study, we developed and evaluated analytic scoring rubrics for two TOEFL CBT writing prompts, investigated the usefulness of analytic scoring in providing diagnostic feedback, and examined the relationships between the analytic scores assigned by human raters and the essay feature variables used in *e-rater*. Even though the analytic scores were highly correlated among themselves and with holistic scores, it was demonstrated that these analytic scores are distinct to some extent and could be used to discriminate among some clusters of ESL learners with different profiles of strengths and weaknesses. We found that some of the *e-rater* essay feature variables for organization, lexical complexity, and mechanics were working as intended in capturing some of the valid rating dimensions valued by human raters in analytic scoring. This provided some support for the validity of the automated scores computed based on these features. Further investigation is warranted to explore effective ways to compute meaningful trait scores and generate useful writing feedback to the ESL learners based on the essay feature variables used in *e-rater* and *Critique*.

In addition, a more conceptual investigation of the relationships between these two kinds of information proved very useful for strengthening the validity evidence for automated scores and also for identifying areas needing further refinement. We have found that there are some important writing features that are not being captured explicitly in *e-rater* and *Critique*. Clearly, some additional variables need to be added to capture the sentence variety/construction aspect of essay quality, the depth of development, coherence, and appropriateness of lexical choice in the ESL learners' essays.

Recommendations and Avenues for Future Research

The ultimate goal of further research of the type done here would be to refine the essay feature variables used in *e-rater* and to create a principled theoretical framework for generating automated trait/analytic scores and diagnostic feedback on the basis of these refined variables. To this end, we suggest the following areas for future research.

Consistency of analytic scores across different writing tasks. Because TOEFL CBT Writing is a single-prompt based writing assessment for each examinee, we were not able to investigate the consistency of analytic scores across different prompts for the same examinees. It is well known that score inconsistency across tasks is a major source of measurement error in performance-based writing assessment. It remains to be seen what analytic rating dimensions are contributing most to such across-task performance inconsistency of examinees and what impact they may have on the examinee clustering based on analytic scores. Analytic scoring of multiple writing prompts taken by the same group of ESL learners is expected to provide data for such investigation.

Creation of e-rater essay features for sentence/syntactic variety. An earlier version of *e-rater* (version 1.3) included various syntactic variables such as counts and ratios of complement, infinitive, relative, and subordinate clauses (Burstein, 2003). It seems that some of these variables can be brought back to create *e-rater* essay feature variables that capture the syntactic variety aspect of essay quality. Further investigation is necessary to explore this possibility. In relation to these, the *Critique* Writing Analysis Tools are continuously updated, as new, enhanced capabilities are added to the tools to detect more ESL-relevant writing errors. These include detecting grammar/usage errors related to the use of articles (Han, Chodorow, & Leacock, 2006), prepositions, and word collocations. It would be interesting to see how the strength of relationships between the *e-rater* linguistic accuracy ratio variables and the analytic scores assigned by human raters changes, as more ESL-relevant errors are reflected into the linguistic accuracy ratio variables used in *e-rater*.

Essay length. Correlation analyses revealed that high correlations among holistic and analytic scores were largely attributable to the impact of essay length on both analytic and holistic scoring. Further research is recommended to explore ways to constrain or otherwise control for essay length, thereby decreasing the influence of this variable and possibly increasing the distinctiveness of analytic scores. Another related research idea is to replicate similar studies on ESL learners’ essays written under untimed testing conditions with a strict essay length requirement imposed for each essay.

References

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Ashwell, T. (2000). Pattern of teacher response to student writing in a multiple draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing*, *9*(3), 227–257.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e-rater*® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved April 16, 2007, from http://www.jtla.org.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29,* 371–383.

Bennett, R. E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS Research Memorandum No. RM-04-01). Princeton, NJ: ETS.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It’s not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–17.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Brennan, R. L. (1999). *Manual for mGENOVA version 2.0.* Iowa City, IA: The University of Iowa.

Burstein, J. (2003). The *e-rater*® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays* (ETS Research Rep. No. RR-98-15). Princeton, NJ: ETS.

Carson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (TOEFL Research Rep. No. 19). Princeton, NJ: ETS.

Chung, G. K W. K., & Baker, E. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23–40). Mahwah, NJ: Lawrence Erlbaum.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston, MA: Heinle & Heinle.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly*, *29*, 762–765.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31–51.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*(1), 67–96.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139–155.

ETS. (1998). *Computer-based TOEFL score user guide*. Princeton, NJ: Author.

Fathman, A., & Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178–190). Cambridge, UK: Cambridge University Press.

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiencies. *TESOL Quarterly*, *28*, 414–420.

Ferris, D. (1995). Student reactions to teacher response in multiple-draft composition classrooms. *TESOL Quarterly, 29*, 33–53.

Ferris, D. (2002). *Treatment of error in second language writing*. Ann Arbor, MI: The University of Michigan Press.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL Test of Written English* (TOEFL Research Rep. No. 64). Princeton, NJ: ETS.

Freedman, S. W. (1984). The registers of student and professional expository writing. Influences on teacher responses. In R. Beach & S. Bridwell (Eds.), *New directions in composition research* (pp. 334–347). New York: Guilford Press.

Gentile, C., Riazantseva, A., & Cline, F. (2002). *A comparison of handwritten and word processed TOEFL essays: Final report.* Unpublished manuscript, ETS, Princeton, NJ.

Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology, 2*(1), 27–38.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123–145.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*, 759–762.

Han, N.-R., Chodorow, M., & Leacock, C. (2006, June). Detecting errors in English article usage by nonnative speakers. *Natural Language Engineering, 12*(2), 115–129.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237–263.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Jarvis, S. (2002). Short texts, best-fitting curves, and new measures of lexical diversity. *Language Testing*, *19*, 57–84.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner composition. *Journal of Second Language Writing*, *12*, 377–403.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3–31.

Laufer, B., & Nation, P. (1995). A vocabulary size test of controlled productive ability. *Language Testing*, *16*, 33–51.

Leacock, C., & Chodorow, M. (2003). Automated grammatical error detection. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195–207). Hillsdale, NJ: Lawrence Erlbaum.

Lee, Y.-W., & Kong, N. (2004). *A preliminary investigation of feature organization frameworks for automated essay scoring and feedback.* Unpublished manuscript, ETS Princeton, NJ.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1998). *A user's guide to FACETS: A Rasch measurement computer program*. Chicago: MESA Press.

McDonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, *30*(1), 105–110.

McNamara, T. (1996). *Measuring second language performance.* New York: Addison Wesley Longman.

Monaghan, W., & Bridgman, B. (2005, April). *E-rater as a quality control on human scores* (R&D Connections No. 2). Princeton, NJ: ETS.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly 24*, 427–442.

Reid, J. (1986).Using the Writer's Workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167–188). Washington, DC: TESOL.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct writing assessments. *Applied Measurement in Education*, *7*(2), 159–170.

Romesburg, H. C. (2004). *Cluster analysis for researchers.* Morrisville, NC: Lulu Press.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative speaking students. *TESOL Quarterly, 22*, 69–90.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, *16*(4), 457–478.

Sheehan, K. S. (2003). *An analysis of e-rater scoring for TOEFL CBT essays.* Unpublished manuscript,: ETS, Princeton, NJ.

Silva, T., & Brice, C. (2004). Research in teaching writing. *Annual Review of Applied Linguistics, 24,* 70–106.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. L., Reed, M., et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement, 59*(3), 492–506.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large scale evaluation of writing. *Research in the Teaching of English, 17,* 285–296.

Weigle, S. C. (2002). *Assessing writing*. New York: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, NJ: Prentice Hall Regents.

Williamson, M. M. (1993). An introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1–43). Cresskill, NJ: Hampton Press.

Williamson, M. M., & Huot, B. A. (Eds.). (1993). *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Young, F. W., & Lewyckyj, R. (1979). *ALSCAL-4 user’s guide.* Carrboro, NC: Data Analysis and Theory Associates.

Notes

[1] At the time this report was written, Yong-Won Lee was on staff at ETS. Currently, Lee is a faculty member of Seoul National University.

[2] The terms *analytic scoring* and *multitrait scoring* are used synonymously in this report.

[3] In the multivariate design, a superscript filled-circle (•) next to a facet symbol indicates that the facet is crossed with the fixed category facet (*v*), whereas a superscript empty circle (o) signals that the facet is nested within the multivariate variables (*v*).

[4] SPSS ALSCAL uses the Euclidian distance model as a basis to compute optimal distances between objects (or items) in an *n*-dimensional space. The Euclidian distance function is derived from the Pythagorean theorem, and is defined as the length of the hypotenuse linking two points in a hypothetical right triangle. The Euclidian distance function is given by $d_{ij}^2 = \sum_a (x_{ia} - x_{ja})^2$,where d_{ij}^2 is the squared Euclidian distance between points *i* and *j*, x_{ia} and x_{ja} are the respective coordinates of points *i* and *j* on axis *a*. See Borg and Groenen (1997) and Giguère (2006) for more information.

[5] The S-Stress is a measure of badness of fit between the hypothesized structure and the original data, which is used in the ASCAL optimization process in MDS analysis. In other words, the S-Stress represents the proportion of variance of the data not accounted for by the MDS model. In the ASCAL optimization process, a series of estimation steps are repeated (or iterated) until the final solution is reached. After each iteration, the current value of S-Stress is compared to the value of S-Stress from the previous iteration. If the improvement is less than a specified value (or converged), iteration stops and the analysis output is generated (see Giguère, 2006 for more details).

[6] It should be noted that measurement facets are conceptualized somewhat differently in MFRM (or Rasch IRT) and generalizability (G) theory. In G-theory, there is a clear distinction made between the object of measurement (usually examinees) and measurement facets, whereas both the object of measurement and measurement facets are all counted as facets in MFRM. Thus, rater-mediated assessment involving examinees, tasks, and raters typically is regarded as a two-facet scenario in G-theory, whereas it is regarded as a three-facet one in MFRM.

[7] In MFRM analysis, there are two different types of fit statistics available: (a) infit (information-weighted, inlier-pattern-sensitive fit statistics) and (b) outfit (outlier-sensitive fit statistics). These two types of statistics can be reported in mean square and standardized values. The mean square value ranges from 0 to infinity, with a expectation of 1. The acceptable range of the mean square values is 0.75 to 1.3 (McNamara, 1996). The mean square value less than 0.75 indicates too little variation or lack of independence, whereas values larger than 1.3 indicate significant misfit or unmodeled excess variation. In this study, the infit mean square values (greater than 1.4) are used as a criterion for identifying misfitting examinees. See Linacre (1998) and McNamara (1996) for more information.

Appendix A

Analytic Scoring Rubrics for TOEFL CBT Writing Prompts

Scoring Rubrics for Development of Ideas

Level 1: Low English Proficiency

Due to problems with English Proficiency, the main points are very difficult to understand. The words may not be in the form of standard English (confusing word order, wrong word forms, frequent misspellings), so that it is hard to know what the writer is trying to say.

Level 2: Limited Response

Due to the limited response given, the development of ideas cannot be judged. Because the response has fewer than eight full typed lines of text (or fewer than 90 words), there is not enough evidence to judge development

Level 3: Minimal Development

Only a few of the main points (less than half) are developed with supporting details, explanations or brief examples. One of the reasons may be developed with a brief example, but the other reasons are simply stated. Or, an explanation is given for part of the writer's opinion, but the rest of the opinion is not developed. Or the essay is mostly a list of ideas supporting the opinion or discussing the theme.

Level 4: Basic Development

Most of the main points (half or more) are developed to one or two levels of depth, although some of the main points may not be developed. One common pattern is: reason → explanation; reason → explanation; reason → explanation. Another is: main point → explanation → example; main point → example; main point → explanation; main point.

Level 5: Some Depth of Development

One of the main ideas is developed in depth, to the third level of development, such as: main ideas → explanation → example → conclusion or main point → problem → solution → example. All of the ideas may not be developed or may only be somewhat developed, but one main idea is developed in depth.

Level 6: More Depth of Development

At least two of the main points are developed in depth (to the third level). For the academic debate essay, a common pattern is: reasons for opinion → explanation → example → conclusion. For the kids and sports essay, a common pattern is idea → explanation → problem → solution.

Scoring Rubric for Organization

Level 1: Low English Proficiency

Due to problems with English proficiency, the points the writer is trying to make are unclear. The words may not be in the form of standard English and/or the words may be so out of order that one cannot understand the essay.

Level 2: Limited Response

Due to the limited response given, the organization of ideas cannot be judged. Because the response has less than eight typed lines of text (or few than 90 words), there is not enough evidence to judge organization

Level 3: Some Organization of Ideas

Some of the ideas flow logically, but most read more like a list of ideas about the topic. In one or two parts of the essay, the writer made some decisions about how to present ideas, i.e., how to order ideas to make a point. However, many of the ideas do not flow logically, the writer changes direction suddenly, interrupting the flow, often making it hard for the reader to understand the main points.

Level 4: More Organization of Ideas

Most of the ideas flow logically (although there still may be a few sudden changes in direction). The writer has clearly made decisions about how to order ideas to make a point, making it easier for the reader to understand the main points.

Level 5: Basic Overall Essay Structure

There is an overall structure to the essay, but it is very basic. The writer may use the prompt to structure the essay (i.e., discussing advantages in one paragraph and disadvantages in another or discussing reasons to support an opinion). OR the structure provided by the thesis statement is not followed. Within the essay, most of the ideas flow logically (although there may be a few sudden changes in direction)

Level 6: Advanced Overall Essay Structure

The overall structure to the essay is very clear and involves an organizing principle or theme that goes beyond the structure of the prompt. Those using the Road Map approach, articulate their organizational structure at the beginning of their essays. Those using the Journey of Discovery approach, articulate their organizational structure near the end of their essays. Within the essay, most of the ideas flow logically (although there may be sudden changes in direction).

Scoring Rubric for Vocabulary

Level 1: Not Enough Evidence

Due to the limited response given, the writer's command of vocabulary cannot be judged. Because the response has less than eight full typed lines of text (or less than 90 words), there is not enough evidence to judge vocabulary.

Level 2: Basic

The essay is mostly comprised of basic words. The range of words is limited to simple expressions, words copied from the prompt, and basic vocabulary that is often used repeatedly. Papers that are longer than eight full lines of text but are difficult to understand are also classified as Basic.

Level 3: Predictable

The essay now includes a mixture of descriptive words and basic words. But most of these words are within a predictable range for students at this level (those taking the TOEFL exam).

Level 4: More Varied

More of the words are descriptive and a wider range of these words is now used. There may be an attempt use more specialized words, but these words are not used correctly.

Level 5: Effective

At this level, all three types of words are used including specialized words, and the range of vocabulary is effective. This represents a more sophisticated control over vocabulary.

* Please note that words provided in the prompt do not count towards the vocabulary rating

* As students begin to use more sophisticated words, they often misspell these new words. The misspellings count as mechanics errors, and the word contributes towards the range of vocabulary.

* However, if they misuse a word, this word does not contribute towards the range of vocabulary.

Scoring Rubric for Sentence Variety and Construction

Level 1: Not Enough Evidence

Due to the limited response given, the writers' command of sentence variety and construction cannot be judged. Because the response has less than eight full typed lines of text (or less than 90 words), there is not enough evidence to judge variety and construction.

Level 2: Minimal Control

Mostly simple sentence structures are used, with little variety; OR almost all of the sentences have the same structure; OR the order of words is so irregular that it is hard to understand the main points.

Level 3: Some Control

There is some variety in sentence structure; some more complex structures are used. However, the attempt to use more complex structures often results in awkwardly constructed sentences.

Level 4: Adequate Control

A wider variety of sentence structures is used. While some of the complex sentences may be awkward, others are well structured.

Level 5: Basic Overall Essay Structure

Writers use a variety of sentence structures to effectively convey the main points. Most of the more complex sentences are well-structured.

Scoring Rubric for Grammar and Usage

Level 1: Not Enough Evidence

Due to the limited response given, the writer's pattern of grammatical errors cannot be judged. Because the response has less than eight full typed lines of text (or less than 90 words), there is not enough evidence to judge the writer's control over usage.

Level 2: Minimal Control

Grammatical errors are constant—75% of the sentences have grammatical errors. OR the grammatical errors are so serious that it is hard to understand the main points.

Level 3: Some Control

There are frequent errors across the paper, but the errors do not interfere with understanding the main points. More than half of the sentences contain grammatical errors (51–74%).

Level 4: Adequate Control

There are not as many grammatical errors across the paper and these errors do not interfere with understanding the main points and subpoints. Half or less than half of the sentences contain grammatical errors (26–50%). Also, the types of errors tend to be aspects of usage that are acquired at later stages of second language development (such as the rules for the use of prepositions and articles).

Level 5: Strong Control

There are few, minor grammatical errors across the paper so that it is easy to understand the main point and subpoints. One-quarter or less than one-quarter of the sentences contain grammatical errors (0–25%). The types of errors are aspects of usage that are acquired at later stages of second language development.

Scoring Rubric for Mechanics

Level 1: Not Enough Evidence

Due to the limited response given, the writer's command of mechanics cannot be judged. Because the response has less than eight full typed lines of text (or less than 90 words), there is not enough evidence to judge mechanics.

Level 2: Minimal Control

Mechanical errors are constant—75% of the sentences have mechanical errors. OR the errors in mechanics are so serious that it is hard to understand the main points. This sometimes happens with frequent spelling and punctuation errors.

Level 3: Some Control

There are frequent errors across the paper, but the errors do not interfere with understanding the main points. More than half of the sentences contain errors in mechanics (51–74%).

Level 4: Adequate Control

There are not as many errors across the paper and the errors do not interfere with the understanding of the main points and most of the subpoints. Half or less than half of the sentences contain errors in mechanics (26–50%).

Level 5: Strong Control

There are few errors across the paper. One-quarter or less than one-quarter of the sentences contain errors in mechanics (0–25%).

Appendix B

Holistic Scoring Rubrics for TOEFL CBT Writing Prompts

The content of this appendix is excerpted from the *Computer-Based TOEFL Test Score User Guide* (ETS, 1998).

6 An essay at this level

- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice, though it may have occasional errors

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and well developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of the language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level

- addresses the writing topic, but slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- displays adequate but possibly inconsistent facility with syntax and use, and it may contain some errors that occasionally obscure meaning

3 An essay at this level my reveal one or more of the following weaknesses

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by one or more of the following weaknesses

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level

- may be incoherent
- may be underdeveloped
- may contain severe and persistent writing errors

0 An essay will be rated 0 if it

- contains no response
- merely copies the topic
- is off-topic, is written in a foreign language, or consists only of keystroke characters

Appendix C

E-rater Essay Feature Variables for Version 2.0

1. Organization and Development

Discourse unit score (DT_OPT) = the difference between the actual and optimal number of discourse units in the essay. The optimal number of discourse units is defined to be 8, which include the introduction, six units for the main body, and the conclusion.

Discourse unit length (DT_AVUL) = the average length of the discourse units in terms of the average number of words across the discourse units.

2. Lexical Complexity

Type/token ratio (CVA_TTVP) = a measure of lexical variety measured in terms of the ratio of types (unique words) over the total number of words among content words in an essay.

Word length (WORDLN) = a measure of average word length in an essay measured in terms of the average number of characters across all words used in the essay.

Vocabulary level (WFLOW): WFLOW is a measure of lexical sophistication that represents vocabulary level of an essay based on standard frequency index.

3. Prompt-Specific Vocabulary Usage (Content Word Vector)

Word vector score (EGW) = a score point value (1–6) for which the maximum cosine correlation over the six score point correlations was obtained. This feature indicates the score point level to which the essay text is most similar with regard to vocabulary usage.

Word vector correlation (COS_EGW) = a cosine correlation value between the essay vocabulary and the sample essays at the highest score point (6).This feature indicates how similar the essay vocabulary is to the vocabulary of the best essays.

4. Errors in Grammar, Usage, Mechanics, and Style

Grammatical accuracy ratio (GRAMMARP) = a grammatical accuracy ratio measure. To obtain the grammatical accuracy ratio measure, the following steps are taken: sum all grammar errors in each essay, divide it by the total number of words, and subtract the result from 1. That is, GRAMMARP = 1 – (number of grammar errors ÷ total number of words).

Usage accuracy ratio (USAGEP) = a usage accuracy ratio measure. To obtain the usage accuracy ratio measure, the following steps are taken: sum all usage errors in each essay, divide it by the total number of words, and subtract the result from 1. That is, USAGEP = 1 – (number of usage errors ÷ total number of words).

Mechanical accuracy ratio. (MECHANIP) = a mechanical accuracy ratio measure. To obtain the mechanical accuracy ratio measure, the following steps are taken: sum all mechanical errors in each essay, divide it by the total number of words, and subtract the result from 1. That is, MECHANIP = 1 – (number of mechanical errors ÷ total number of words).

Stylistic accuracy ratio (STYLEP) = a measure of stylistic accuracy ratio. To compute this variable for each essay, the following steps are taken: sum all stylistic errors in an essay, divide it by the total number of words in the essay, and take a minus log of the result. That is, STYLEP = [1 – (number of style errors ÷ total number of words)].

5. Essay Length

Essay length (TNW) = the total number of words in each essay.

Appendix D

One- and Three-Dimensional Plots of Holistic and Analytic Scores Based on Multidimensional Scaling

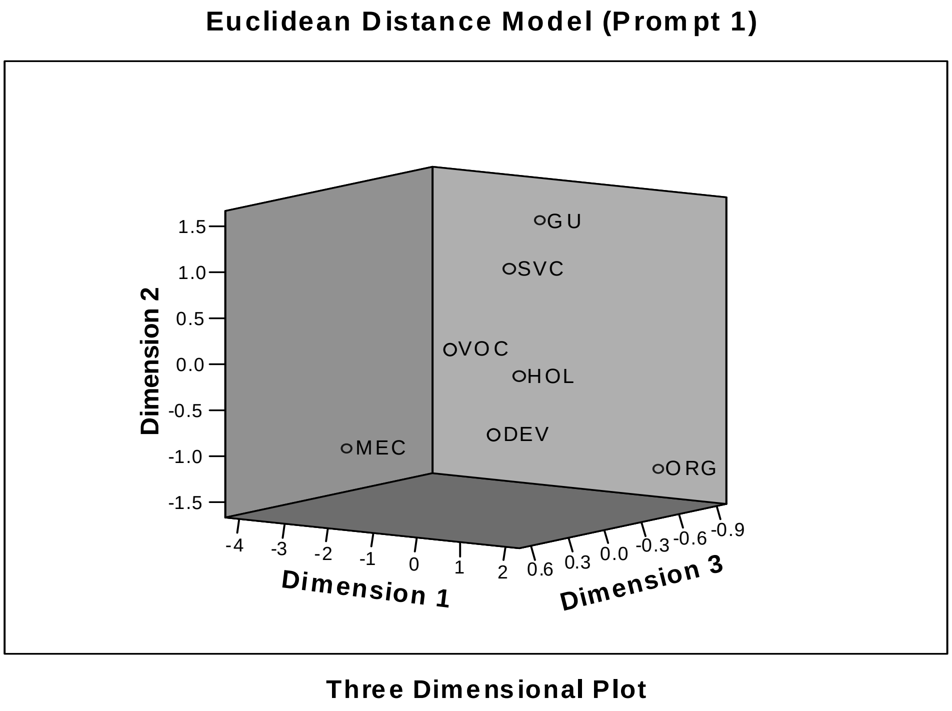

Figure D1. **Euclidean distance model (Prompt 1): Three-dimensional plot.**

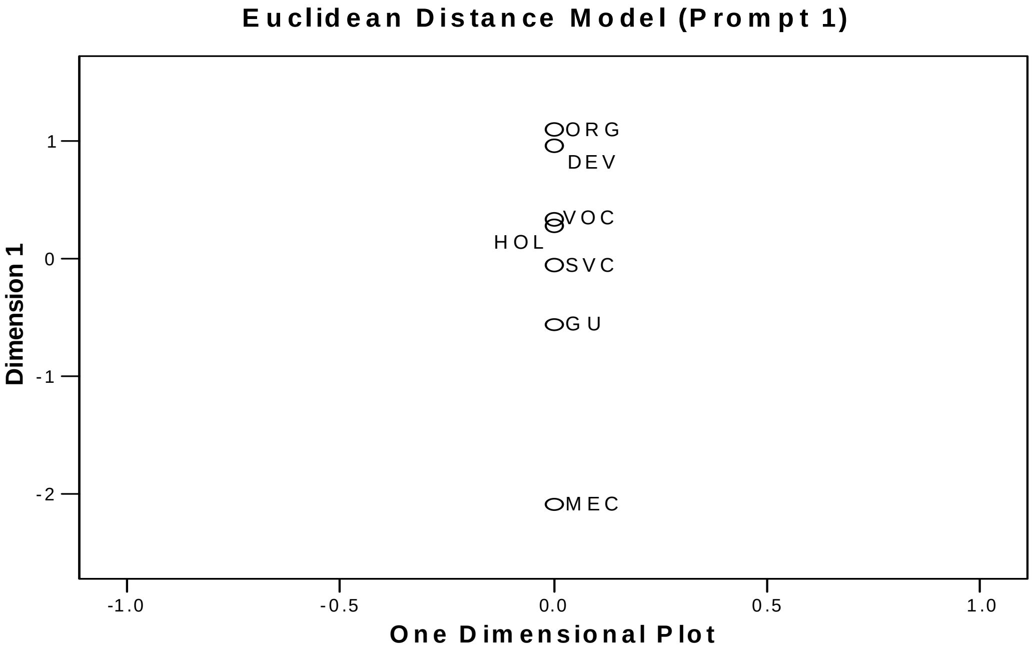

Figure D2. **Euclidean distance model (Prompt 1): One-dimensional plot.**

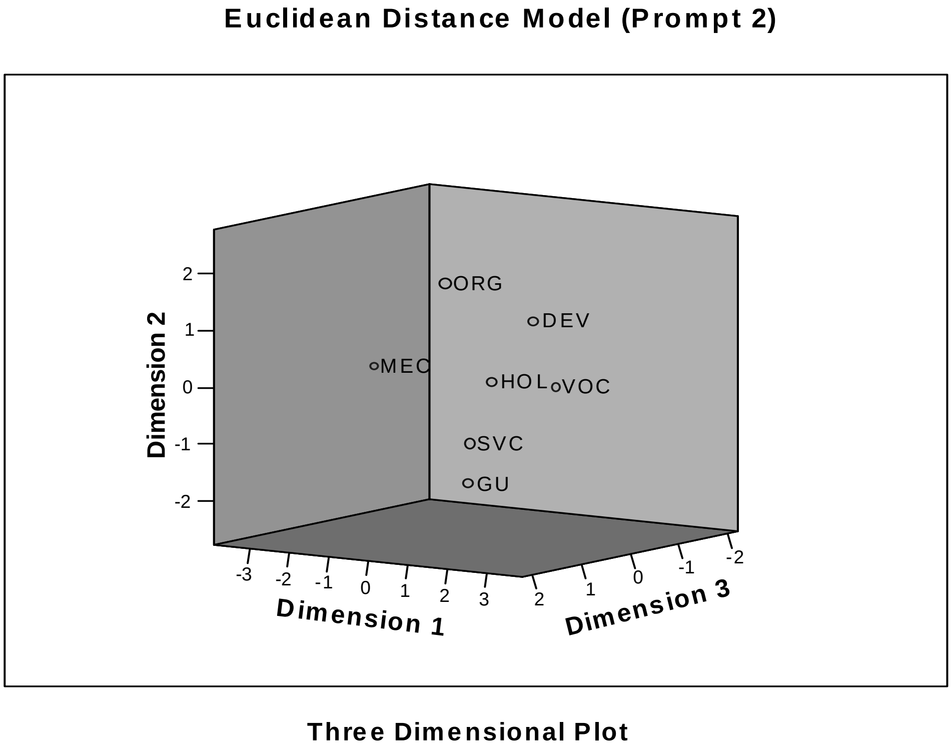

Figure D3. **Euclidean distance model (Prompt 2): Three-dimensional plot.**

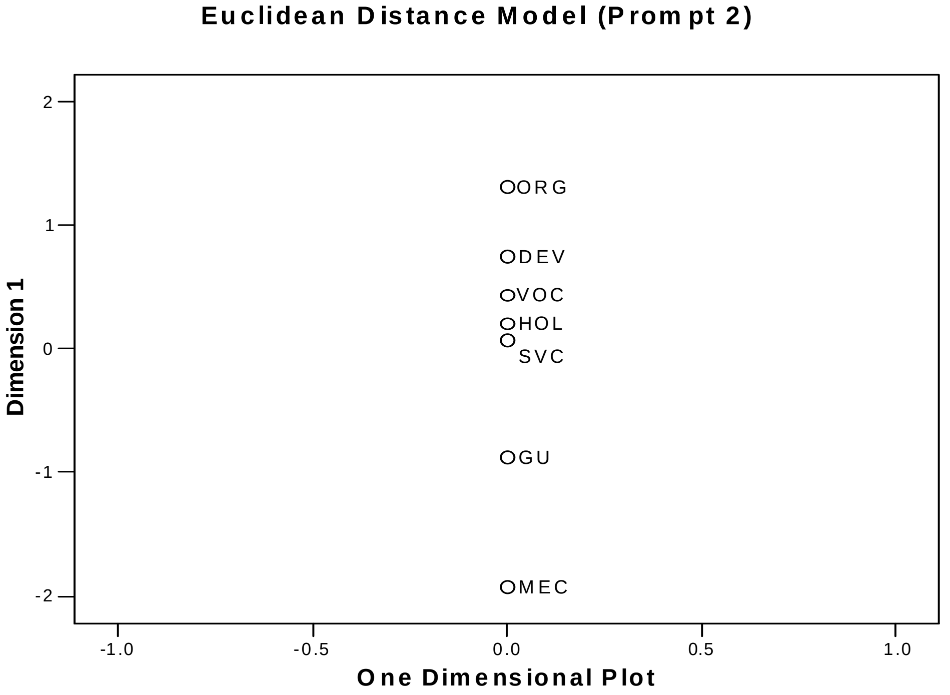

Figure D4. **Euclidean distance model (Prompt 2): One-dimensional plot.**

Appendix E

Score Profiles for Examinee Clusters Obtained from Cluster Analyses of Six Analytic Scores for the Two Prompts

Table E1

Score Profiles for Examinee Clusters Obtained From Cluster Analyses of Six Analytic Scores for Prompt 1

ID	Cluster	Raw analytic scores						Standardized analytic scores						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEW	ORG	VOC	SVC	GU	MEC	Raw	SD		
								Cluster 1									
370	0.9	3.0	3.5	2.5	1.5	1.5	4.0	-0.5	-0.3	-0.5	-1.2	-1.1	0.7	3.0	-0.5	132	JPN
136	1.0	2.0	3.0	2.0	2.0	2.5	4.0	-1.4	-0.7	-0.9	-0.8	-0.3	0.7	2.0	-1.2	112	CHI
46	1.1	2.0	2.5	2.5	2.0	2.0	4.0	-1.4	-1.1	-0.5	-0.8	-0.7	0.7	3.0	-0.5	137	KOR
287	1.1	3.0	3.0	2.5	2.0	2.0	5.0	-0.5	-0.7	-0.5	-0.8	-0.7	1.5	3.0	-0.5	119	JPN
83	1.2	3.0	3.5	2.0	2.0	2.5	5.0	-0.5	-0.3	-0.9	-0.8	-0.3	1.5	2.0	-1.2	130	KOR
277	1.3	3.0	3.5	2.0	1.5	1.5	3.0	-0.5	-0.3	-0.9	-1.2	-1.1	-0.1	3.0	-0.5	139	JPN
322	1.4	3.0	2.5	2.0	1.5	1.5	3.0	-0.5	-1.1	-0.9	-1.2	-1.1	-0.1	2.0	-1.2	134	CHI
102	1.5	2.5	2.5	3.0	3.0	2.0	4.5	-0.9	-1.1	-0.1	0.0	-0.7	1.1	3.5	-0.1	166	JPN
265	1.6	3.0	2.5	2.0	3.0	3.0	3.5	-0.5	-1.1	-0.9	0.0	0.1	0.3	2.5	-0.8	123	CHI
330	1.6	3.0	3.5	3.0	1.5	1.5	5.0	-0.5	-0.3	-0.1	-1.2	-1.1	1.5	3.0	-0.5	119	JPN
353	1.6	3.0	3.0	2.0	1.0	1.0	3.5	-0.5	-0.7	-0.9	-1.6	-1.5	0.3	2.0	-1.2	103	SPA
329	1.7	3.0	3.0	2.5	3.0	3.0	5.0	-0.5	-0.7	-0.5	0.0	0.1	1.5	3.5	-0.1	129	JPN

(Table continues)

Table E1 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic scores						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEW	ORG	VOC	SVC	GU	MEC	Raw	SD		
268	1.8	3.0	2.5	2.5	2.0	3.5	3.0	-0.5	-1.1	-0.5	-0.8	0.6	-0.1	3.0	-0.5	169	SPA
197	1.9	3.0	3.0	1.5	3.0	3.5	4.0	-0.5	-0.7	-1.3	0.0	0.6	0.7	3.0	-0.5	118	JPN
340	2.0	2.0	3.5	1.5	1.0	1.0	4.5	-1.4	-0.3	-1.3	-1.6	-1.5	1.1	2.5	-0.8	99	JPN
443	2.4	3.5	2.5	2.5	3.0	3.5	2.5	-0.1	-1.1	-0.5	0.0	0.6	-0.5	2.5	-0.8	112	JPN
334	2.4	3.0	4.5	3.5	1.0	1.0	4.5	-0.5	0.6	0.3	-1.6	-1.5	1.1	3.0	-0.5	112	KOR
230	2.5	3.5	5.0	2.0	2.5	2.0	2.5	-0.1	1.0	-0.9	-0.4	-0.7	-0.5	3.0	-0.5	200	SPA
Mean		2.9	3.2	2.3	2.0	2.1	3.9	-0.6	-0.5	-0.6	-0.8	-0.6	0.7	2.8	-0.6	131	
Cluster 2																	
161	0.8	4.5	6.0	3.5	3.5	2.5	3.5	0.8	1.8	0.3	0.3	-0.3	0.3	4.0	0.3	281	ITA
306	0.9	5.0	5.5	4.0	3.5	3.0	2.5	1.2	1.4	0.7	0.3	0.1	-0.5	5.5	1.4	350	TEL
415	1.1	5.0	6.0	4.0	4.0	3.0	3.0	1.2	1.8	0.7	0.7	0.1	-0.1	4.5	0.6	338	ENG
303	1.1	5.0	6.0	3.0	3.5	2.0	3.0	1.2	1.8	-0.1	0.3	-0.7	-0.1	4.0	0.3	296	KOR
269	1.5	5.5	5.0	4.5	2.5	3.0	2.5	1.7	1.0	1.1	-0.4	0.1	-0.5	5.0	1.0	373	SPA
384	1.5	5.5	6.0	4.0	2.5	3.5	4.0	1.7	1.8	0.7	-0.4	0.6	0.7	4.5	0.6	261	KOR
215	1.5	4.5	6.0	3.0	4.0	2.5	4.0	0.8	1.8	-0.1	0.7	-0.3	0.7	5.0	1.0	324	JPN
364	1.6	4.5	5.0	4.5	3.5	2.5	2.0	0.8	1.0	1.1	0.3	-0.3	-0.9	4.5	0.6	316	*
428	1.9	6.0	5.5	4.5	2.0	2.5	4.0	2.1	1.4	1.1	-0.8	-0.3	0.7	6.0	1.7	446	BEN

(Table continues)

Table E1 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic scores						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEW	ORG	VOC	SVC	GU	MEC	Raw	SD		
266	1.9	4.0	5.0	2.5	2.5	2.0	3.5	0.4	1.0	-0.5	-0.4	-0.7	0.3	4.0	0.3	198	JPN
Mean		5.0	5.6	3.8	3.2	2.7	3.2	1.2	1.5	0.5	0.1	-0.1	0.1	4.7	0.8	318	
Cluster 3																	
201	0.7	1.5	2.0	2.0	2.0	2.0	2.0	-1.8	-1.5	-0.9	-0.8	-0.7	-0.9	1.5	-1.5	100	*
403	1.0	2.0	1.5	2.0	2.0	1.5	1.0	-1.4	-1.9	-0.9	-0.8	-1.1	-1.7	2.0	-1.2	97	JPN
309	1.2	3.0	1.0	2.0	2.0	2.0	2.0	-0.5	-2.3	-0.9	-0.8	-0.7	-0.9	2.0	-1.2	151	DUT
242	1.4	2.5	2.0	2.0	3.0	2.0	1.5	-0.9	-1.5	-0.9	0.0	-0.7	-1.3	2.0	-1.2	95	ARA
460	1.6	1.0	1.5	2.0	2.5	2.5	1.5	-2.2	-1.9	-0.9	-0.4	-0.3	-1.3	1.0	-1.9	111	KOR
407	1.7	3.0	1.0	2.0	2.5	2.0	3.0	-0.5	-2.3	-0.9	-0.4	-0.7	-0.1	2.0	-1.2	129	KOR
439	1.8	2.0	2.0	3.0	1.0	1.0	1.0	-1.4	-1.5	-0.1	-1.6	-1.5	-1.7	1.0	-1.9	29	JPN
137	1.8	2.5	3.0	3.0	1.5	2.0	1.5	-0.9	-0.7	-0.1	-1.2	-0.7	-1.3	2.5	-0.8	110	SPA
171	1.8	2.0	2.0	1.0	1.0	1.0	3.0	-1.4	-1.5	-1.7	-1.6	-1.5	-0.1	3.0	-0.5	88	JPN
114	2.2	1.0	1.0	1.0	1.0	1.0	3.0	-2.2	-2.3	-1.7	-1.6	-1.5	-0.1	1.5	-1.5	69	*
Mean		2.1	1.7	2.0	1.9	1.7	2.0	-1.3	-1.7	-0.9	-0.9	-0.9	-0.9	1.9	-1.3	98	
Cluster 4																	
25	1.4	4.0	3.5	3.5	4.0	5.0	5.0	0.4	-0.3	0.3	0.7	1.8	1.5	4.0	0.3	165	ENG
75	1.4	3.5	4.5	3.0	5.0	5.0	5.0	-0.1	0.6	-0.1	1.5	1.8	1.5	3.0	-0.5	143	999

(Table continues)

Table E1 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic scores						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEW	ORG	VOC	SVC	GU	MEC	Raw	SD		
81	1.4	3.0	3.5	4.0	4.0	4.0	5.0	-0.5	-0.3	0.7	0.7	1.0	1.5	4.0	0.3	162	OTH
66	1.6	4.0	5.5	4.5	3.5	5.0	5.0	0.4	1.4	1.1	0.3	1.8	1.5	6.0	1.7	285	FRE
56	1.7	4.0	3.5	5.0	4.5	4.0	4.5	0.4	-0.3	1.6	1.1	1.0	1.1	4.0	0.3	211	POR
157	1.8	3.5	3.5	3.0	5.0	4.0	4.0	-0.1	-0.3	-0.1	1.5	1.0	0.7	4.5	0.6	219	GER
317	1.8	4.5	6.0	4.0	4.5	3.5	5.0	0.8	1.8	0.7	1.1	0.6	1.5	4.5	0.6	240	KOR
101	2.0	4.5	6.0	4.5	5.0	3.5	4.5	0.8	1.8	1.1	1.5	0.6	1.1	5.5	1.4	371	POL
267	2.2	4.0	5.0	2.5	2.5	4.0	5.0	0.4	1.0	-0.5	-0.4	1.0	1.5	4.0	0.3	219	JPN
Mean		3.9	4.6	3.8	4.2	4.2	4.8	0.3	0.6	0.6	0.9	1.2	1.3	4.4	0.6	224	
Cluster 5																	
140	0.5	5.5	5.5	4.5	4.5	4.0	2.5	1.7	1.4	1.1	1.1	1.0	-0.5	5.0	1.0	341	CHI
447	1.0	6.0	5.5	4.0	4.5	4.0	2.5	2.1	1.4	0.7	1.1	1.0	-0.5	6.0	1.7	423	ENG
440	1.0	6.0	5.0	5.0	4.5	4.0	3.0	2.1	1.0	1.6	1.1	1.0	-0.1	5.0	1.0	291	GEN
435	1.2	6.0	5.5	5.0	5.0	4.5	2.5	2.1	1.4	1.6	1.5	1.4	-0.5	5.0	1.0	316	SWE
341	1.2	5.0	4.5	4.0	4.0	3.5	2.0	1.2	0.6	0.7	0.7	0.6	-0.9	5.0	1.0	407	TEL
376	1.4	4.5	5.0	4.0	5.0	5.0	2.5	0.8	1.0	0.7	1.5	1.8	-0.5	5.5	1.4	316	HIN
459	1.4	6.0	5.5	4.5	5.0	4.5	3.5	2.1	1.4	1.1	1.5	1.4	0.3	6.0	1.7	327	HEB
350	1.4	6.0	6.0	5.0	4.5	4.5	3.0	2.1	1.8	1.6	1.1	1.4	-0.1	6.0	1.7	469	MAR

(Table continues)

Table E1 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic scores						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEW	ORG	VOC	SVC	GU	MEC	Raw	SD		
355	1.5	5.5	4.5	5.0	4.5	3.0	2.0	1.7	0.6	1.6	1.1	0.1	-0.9	3.0	-0.5	444	RUM
260	1.6	5.0	5.0	3.0	4.5	3.5	2.5	1.2	1.0	-0.1	1.1	0.6	-0.5	5.5	1.4	370	SPA
36	1.7	4.0	4.5	5.0	5.0	3.5	3.0	0.4	0.6	1.6	1.5	0.6	-0.1	5.5	1.4	388	ITA
456	1.8	6.0	6.0	4.5	5.0	5.0	3.5	2.1	1.8	1.1	1.5	1.8	0.3	5.5	1.4	383	URD
170	1.9	4.0	5.0	3.5	4.5	4.0	1.5	0.4	1.0	0.3	1.1	1.0	-1.3	4.0	0.3	369	TEL
76	2.6	4.0	3.5	5.0	3.0	4.0	2.5	0.4	-0.3	1.6	0.0	1.0	-0.5	5.5	1.4	356	MAR
Mean		5.3	5.1	4.4	4.5	4.1	2.6	1.4	1.0	1.1	1.2	1.1	-0.4	5.2	1.1	371	

Note. Dist = distribution, DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, MEC = mechanics, TNW = total number of words, NL = nationality, ARA = Arabic, BEN = Bengali, CHI = Chinese, DUT = Dutch, ENG = English, FAS = Farsi, FRE = French, GER = German, GUJ = Gujarati, HEB = Hebrew, HIN = Hindi, IND = Indonesian, ITA = Italian, JPN = Japanese, KOR = Korean, LIT = Lithuanian, MAR = Marathi, POL = Polish, POR = Portuguese, RUM = Romanian, SIN = Sinalese, OTH = Siswathi, SPA = Spanish, SWE = Swedish, TEL = Telugu, THA = Thai, URD = Urdu, * = unknown.

Table E2

Score Profiles for Examinee Clusters Obtained from Cluster Analyses of Six Analytic Scores for Prompt 2

ID	Cluster	Raw analytic scores						Standardized analytic score						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEV	ORG	VOC	SVC	GU	MEC	Raw	SD		
								Cluster 1									
1029	0.7	3.0	4.0	4.0	4.5	3.0	2.0	-0.6	0.1	0.8	1.1	0.2	-1.0	4.0	0.3	185	LIT
1016	1.1	3.0	3.5	4.0	4.5	2.5	3.0	-0.6	-0.3	0.8	1.1	-0.2	-0.2	4.0	0.3	174	ARA
1404	1.1	3.5	4.5	3.0	4.0	3.5	2.0	-0.1	0.6	0.0	0.7	0.6	-1.0	4.5	0.6	285	SWE
1036	1.2	3.5	4.5	3.5	5.0	2.5	2.0	-0.1	0.6	0.4	1.5	-0.2	-1.0	4.0	0.3	234	ARA
1356	1.2	3.5	4.0	3.5	4.0	4.0	2.0	-0.1	0.1	0.4	0.7	1.0	-1.0	4.0	0.3	163	FAS
1118	1.4	3.5	4.5	4.0	3.5	2.0	2.0	-0.1	0.6	0.8	0.3	-0.7	-1.0	4.0	0.3	415	SPA
1431	2.0	3.0	3.0	2.5	3.0	3.0	1.5	-0.6	-0.8	-0.4	-0.1	0.2	-1.4	3.0	-0.5	128	JPN
1085	2.3	3.0	3.5	4.0	4.5	2.5	4.5	-0.6	-0.3	0.8	1.1	-0.2	1.0	4.0	0.3	201	JPN
Mean		3.3	3.9	3.6	4.1	2.9	2.4	-0.4	0.1	0.4	0.8	0.1	-0.7	3.9	0.2	223	
								Cluster 2									
1309	1.2	3.0	1.5	2.0	1.5	1.5	2.0	-0.6	-2.1	-0.9	-1.2	-1.1	-1.0	2.0	-1.2	106	CHI
1265	1.4	3.0	3.0	1.5	1.5	1.5	3.0	-0.6	-0.8	-1.3	-1.2	-1.1	-0.2	2.0	-1.2	102	JPN
1320	1.4	2.0	1.5	1.0	1.0	1.0	2.0	-1.5	-2.1	-1.7	-1.6	-1.5	-1.0	2.0	-1.2	78	JPN
1451	1.4	2.5	2.0	1.5	1.0	1.0	3.0	-1.1	-1.7	-1.3	-1.6	-1.5	-0.2	2.0	-1.2	91	KOR

(Table continues)

Table E2 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic score						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEV	ORG	VOC	SVC	GU	MEC	Raw	SD		
1232	1.4	2.5	2.5	1.0	3.0	2.0	2.5	-1.1	-1.2	-1.7	-0.1	-0.7	-0.6	2.5	-0.8	100	KOR
1415	1.6	1.5	2.0	1.0	2.5	1.5	1.0	-2.0	-1.7	-1.7	-0.5	-1.1	-1.8	1.0	-1.9	28	*
1392	2.3	3.0	4.0	2.5	1.0	1.0	2.0	-0.6	0.1	-0.4	-1.6	-1.5	-1.0	2.0	-1.2	107	KOR
1435	2.4	1.5	2.0	1.0	3.0	3.0	1.0	-2.0	-1.7	-1.7	-0.1	0.2	-1.8	1.0	-1.9	58	JPN
Mean		2.4	2.3	1.4	1.8	1.6	2.1	-1.2	-1.4	-1.3	-1.0	-1.0	-1.0	1.8	-1.3	84	
Cluster 3																	
1119	0.9	3.5	3.5	2.5	2.5	2.5	5.0	-0.1	-0.3	-0.4	-0.5	-0.2	1.4	4.0	0.3	288	JPN
1292	1.0	3.0	4.0	2.0	2.0	1.5	4.5	-0.6	0.1	-0.9	-0.9	-1.1	1.0	3.0	-0.5	124	JPN
1372	1.1	3.0	3.5	2.5	1.5	1.5	4.0	-0.6	-0.3	-0.4	-1.2	-1.1	0.6	3.0	-0.5	120	JPN
1138	1.1	2.5	3.0	3.0	2.0	2.5	4.0	-1.1	-0.8	0.0	-0.9	-0.2	0.6	3.0	-0.5	109	KOR
1383	1.3	3.0	3.0	3.0	1.5	1.5	4.5	-0.6	-0.8	0.0	-1.2	-1.1	1.0	2.0	-1.2	122	JPN
1202	1.4	3.5	2.5	1.5	2.5	2.0	4.5	-0.1	-1.2	-1.3	-0.5	-0.7	1.0	3.0	-0.5	118	JPN
1183	1.4	3.0	3.0	1.5	1.5	1.5	4.0	-0.6	-0.8	-1.3	-1.2	-1.1	0.6	2.5	-0.8	105	KOR
1044	1.5	2.5	2.5	3.0	2.0	2.0	3.5	-1.1	-1.2	0.0	-0.9	-0.7	0.2	3.0	-0.5	106	JPN
1091	1.6	4.0	3.5	2.5	3.0	3.0	5.0	0.3	-0.3	-0.4	-0.1	0.2	1.4	4.0	0.3	263	THA
1342	1.6	4.0	3.5	3.5	2.5	2.5	5.0	0.3	-0.3	0.4	-0.5	-0.2	1.4	3.0	-0.5	165	CHI
1092	1.7	2.5	3.0	2.5	2.5	3.5	5.0	-1.1	-0.8	-0.4	-0.5	0.6	1.4	3.0	-0.5	126	JPN

(Table continues)

Table E2 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic score						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEV	ORG	VOC	SVC	GU	MEC	Raw	SD		
1333	1.7	3.0	4.5	2.0	1.5	1.5	3.5	-0.6	0.6	-0.9	-1.2	-1.1	0.2	2.5	-0.8	133	*
1256	1.9	3.5	2.5	1.5	3.5	2.5	4.0	-0.1	-1.2	-1.3	0.3	-0.2	0.6	3.0	-0.5	145	JPN
1340	2.1	3.0	4.0	3.0	1.0	1.0	3.5	-0.6	0.1	0.0	-1.6	-1.5	0.2	3.0	-0.5	123	JPN
1120	2.8	3.5	5.0	2.0	4.0	3.5	4.5	-0.1	1.0	-0.9	0.7	0.6	1.0	4.0	0.3	245	JPN
Mean		3.2	3.4	2.4	2.2	2.2	4.3	-0.4	-0.4	-0.5	-0.7	-0.5	0.8	3.1	-0.4	153	
Cluster 4																	
1022	1.2	5.5	5.5	5.0	3.0	2.5	4.5	1.7	1.5	1.6	-0.1	-0.2	1.0	5.5	1.4	459	*
1131	1.3	6.0	5.5	4.0	3.5	2.5	4.5	2.2	1.5	0.8	0.3	-0.2	1.0	5.0	1.0	408	KOR
1460	1.4	4.0	5.0	4.0	3.5	2.0	5.0	0.3	1.0	0.8	0.3	-0.7	1.4	3.5	-0.1	216	*
1083	1.4	3.5	6.0	4.0	3.5	2.5	4.0	-0.1	1.9	0.8	0.3	-0.2	0.6	4.0	0.3	240	THA
1461	1.7	4.0	5.5	3.0	2.5	2.0	4.0	0.3	1.5	0.0	-0.5	-0.7	0.6	4.0	0.3	237	KOR
1090	1.8	5.5	6.0	5.0	4.5	3.0	3.5	1.7	1.9	1.6	1.1	0.2	0.2	5.0	1.0	427	*
1124	1.9	3.5	6.0	3.5	3.5	2.5	3.0	-0.1	1.9	0.4	0.3	-0.2	-0.2	4.0	0.3	255	THA
1452	2.0	6.0	6.0	3.5	4.5	2.5	5.0	2.2	1.9	0.4	1.1	-0.2	1.4	5.5	1.4	516	KOR
1062	2.1	4.0	5.0	4.5	2.0	2.0	3.0	0.3	1.0	1.2	-0.9	-0.7	-0.2	5.0	1.0	436	IND
1456	2.2	6.0	4.5	5.0	3.5	3.5	5.0	2.2	0.6	1.6	0.3	0.6	1.4	6.0	1.7	378	TEL
Mean		4.8	5.5	4.2	3.4	2.5	4.2	1.1	1.5	0.9	0.3	-0.2	0.7	4.8	0.8	357	

(Table continues)

Table E2 (continued)

ID	Cluster	Raw analytic scores						Standardized analytic score						Holistic score		TNW	NL
	Dist	DEV	ORG	VOC	SVC	GU	MEC	DEV	ORG	VOC	SVC	GU	MEC	Raw	SD		
Cluster 5																	
1331	0.8	5.0	4.5	3.5	4.5	5.0	3.5	1.3	0.6	0.4	1.1	1.9	0.2	5.0	1.0	352	GER
1189	1.2	5.5	5.5	4.5	5.0	5.0	3.5	1.7	1.5	1.2	1.5	1.9	0.2	5.5	1.4	379	SIN
1380	1.2	4.0	4.5	4.0	4.5	5.0	2.5	0.3	0.6	0.8	1.1	1.9	-0.6	4.5	0.6	283	GUJ
1061	1.3	4.5	5.5	5.0	5.0	4.5	3.0	0.8	1.5	1.6	1.5	1.4	-0.2	6.0	1.7	340	SPA
1377	1.4	5.0	5.0	4.0	4.0	4.0	2.5	1.3	1.0	0.8	0.7	1.0	-0.6	4.5	0.6	498	ARA
1367	1.5	4.0	5.0	5.0	5.0	4.0	3.0	0.3	1.0	1.6	1.5	1.0	-0.2	5.0	1.0	292	BEN
1057	1.5	4.0	3.5	4.0	5.0	5.0	3.5	0.3	-0.3	0.8	1.5	1.9	0.2	5.5	1.4	229	HIN
1319	1.6	6.0	4.5	5.0	5.0	5.0	3.5	2.2	0.6	1.6	1.5	1.9	0.2	6.0	1.7	521	ENG
1464	1.6	4.5	5.5	3.5	5.0	5.0	4.5	0.8	1.5	0.4	1.5	1.9	1.0	5.0	1.0	262	KOR
1035	2.4	5.0	3.5	3.5	3.0	4.5	4.5	1.3	-0.3	0.4	-0.1	1.4	1.0	4.0	0.3	294	GER
Mean		4.8	4.7	4.2	4.6	4.7	3.4	1.0	0.7	1.0	1.2	1.6	0.1	5.1	1.1	345	

Note. Dist = Distribution, DEV = development, ORG = organization, VOC = vocabulary, SVC = sentence variety/construction, GU =grammar/usage, MEC = mechanics, TNW = total number of words, NL = nationality, ARA = Arabic, BEN = Bengali, CHI = Chinese, DUT = Dutch, ENG = English, FAS = Farsi, FRE = French, GER = German, GUJ = Gujarati, HEB = Hebrew, HIN = Hindi, IND = Indonesian, ITA = Italian, LIT = Lithuanian, JPN = Japanese, KOR = Korean, LIT = Lithuanian, MAR = Marathi, POL = Polish, POR = Portuguese, RUM = Romanian, SIN = Sinalese, OTH = Siswathi, SPA = Spanish, SWE = Swedish, TEL = Telugu, THA = Thai, URD = Urdu, AND * = unknown.

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands