ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–15

Validation of Automated Scores of TOEFL iBT® Tasks Against Nontest Indicators of Writing Ability

Sara Cushing Weigle

June 2011

Validation of Automated Scores of TOEFL iBT® Tasks Against Nontest Indicators of Writing Ability

Sara Cushing Weigle
Georgia State University, Atlanta

RR-11-24

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

Automated scoring has the potential to dramatically reduce the time and costs associated with the assessment of complex skills such as writing, but its use must be validated against a variety of criteria for it to be accepted by test users and stakeholders. This study addresses two validity-related issues regarding the use of e-rater® with the independent writing task on the TOEFL iBT® (Internet-based test). First, relationships between automated scores of iBT tasks and nontest indicators of writing ability were examined. This was followed by exploration of prompt-related differences in automated scores of essays written by the same examinees. Correlations between both human and e-rater scores and nontest indicators were moderate but consistent, with few differences between e-rater and human rater scores. E-rater was more consistent across prompts than individual human raters, although there were differences in scores across prompts for the individual features used to generate total e-rater scores.

Key words: automated scoring, writing assessment, second language, validity, e-rater

The TOEFL® exam 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 (2010-2011) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Carol A. Chapelle	Iowa State University
Barbara Hoekje	Drexel University
Ari Huhta	University of Jyväskylä, Finland
John M. Norris	University of Hawaii at Manoa
James Purpura	Columbia University
Carsten Roever	University of Melbourne
Steve Ross	University of Maryland
Mikyuki Sasaki	Nagoya Gakuin University
Norbert Schmitt	University of Nottingham
Robert Schoonen	University of Amsterdam
Ling Shi	University of British Columbia

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® Committee of Examiners and the TOEFL program at ETS. Additional research support came from the Department of Applied Linguistics & ESL at Georgia State University. Yanbin Lu and Amanda Baker assisted with data collection and analysis, and Liang Guo helped with the preparation of the final report. Patricia Carrell served as consultant to the project, primarily assisting with instrument design and planning data collection. ETS provided the essay prompts for the study, scored all the essays with e-rater®, and arranged for human raters for about half of the essays. I would like to thank the site coordinators at the institutions where I collected data: Robert Nelson, Lily Compton, Kristin Di Gennaro, Cameron Jaynes, Jennifer Lund, Megan Kilbourn, Nur Yigitoglu, Dudley Reynolds, Sunyoung Shin, and Youngsoon So. Susan Firestone, Joe Lee, Amanda Baker, Holly Joseph, Caroline Payant, Jason Litzenberg, and Magdi Kandil served as raters for the TOEFL iBT® essays and the submitted writing samples. Finally I would like to thank the reviewers at ETS, and in particular Mary Enright, for their helpful comments on earlier drafts of this report. The responsibility for any remaining errors is solely mine, and the ideas and opinions expressed in the report are those of the author, not necessarily of ETS or the TOEFL program.

Table of Contents

List of Tables

Automated scoring has the potential to reduce dramatically the time and resources associated with the assessment of complex skills such as writing—in particular, the recruitment, training, and monitoring of raters. Much of the research on automated scoring has compared automated scores on essays to the scores given by human raters to the same essays, and has demonstrated convincingly that automated scores are at least as reliable as human scores. However, the increased use of automated scores for both high- and low-stakes testing has sparked a great deal of controversy, particularly among writing teachers, who have expressed a variety of validity-related concerns regarding automated scoring systems. In order for the use of automated scores to be accepted by test users and other stakeholders, empirical research into the meaning of automated scores is crucial.

One test for which automated scoring has recently begun to be used is the TOEFL iBT®, which is required for admission of nonnative speakers of English to many colleges and universities in North America. The TOEFL iBT writing section includes two types of writing tasks: (a) an independent task, in which the test takers are asked to express and support an opinion on a familiar topic, and (b) an integrated task, in which the test takers are required to demonstrate an understanding of the relationship between information in a lecture and a reading. Since July 2009 ETS's automated scoring system e-rater® has been used as one of two raters to score the independent task (see Enright & Quinlan, 2008, for an evaluation of the use of e-rater with this task).

Current approaches to the investigation of validity (Kane, 1992; 2001; Mislevy, Steinberg, & Almond, 2002, 2003) require articulating an interpretive argument for validity by making explicit the chain of inferences that link a test to its use. These inferences begin with defining the domain of interest and end with using test scores to make decision. Each inference is then examined through the collection of evidence that either supports or refutes the inference. Taking this approach, Chapelle, Enright, and Jamieson (2008) provided a framework for investigating for the validity of the TOEFL. This framework includes six inferences that must be supported through empirical studies: domain description, evaluation, generalization, explanation, extrapolation, and utilization. An accumulation of evidence supporting each of these inferences thus supports the interpretive argument for TOEFL validity.

This study addresses two validity-related issues regarding the use of e-rater with the TOEFL iBT: (a) relationships between automated and human scores of TOEFL iBT independent

writing tasks and nontest indicators of writing ability and (b) prompt-related differences in automated scores of essays written by the same examinees. To situate this study within the framework of the TOEFL interpretive argument, the research questions address the inferences of generalization: "observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters" (Chapelle et al., 2008, p. 19) and extrapolation: "performance on the test is related to other criteria of language proficiency in the academic context" (Chapelle et al., 2008, p. 21).

Literature Review

Automated scoring of essays has been possible since the 1960s (Page, 1966) but has only recently been used on a large scale. Several automated scoring systems have been developed, including the PEG system (Page, 2003); the Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2003), and IntelliMetric (Elliot, 2003). This study focuses on e-rater, developed by ETS. For some years e-rater was used operationally, along with human raters, on the Graduate Management Admissions Test (GMAT), and e-rater is also used in ETS' online essay evaluation service, known as Criterion® (Burstein, Chodorow, & Leacock, 2003). This literature review begins with a description of e-rater, followed by a discussion of validity issues related to automated scoring in general and then specifically in terms of evaluating the writing of nonnative speakers of English. The literature review closes with a discussion of the issue of task variability in writing assessment and finally addresses these issues with regard to the validity argument for the TOEFL.

E-rater uses a corpus-based approach to analyze essay characteristics, and is trained on a large set of essays written on a specific prompt to extract a small set of features that are predictive of scores given by human raters. As described by Chodorow and Burstein (2004), these features are generally of four types: syntactic, discourse, topical, and lexical. In earlier versions of e-rater, stepwise linear regression was used to select features for each prompt from the training essays, and these features were used to predict scores given by human raters in cross-validation studies on another set of essays on the same prompt.

The current version of e-rater uses a standard set of features across prompts, allowing both general and prompt-specific modeling for scoring (Attali & Burstein, 2006; Enright & Quinlan, 2010). These features are typically the following, although they may vary for specific analyses:

- *Errors in grammar, usage, mechanics, style*: These errors are extracted from by the writing analysis tools used in the Criterion online essay evaluation service (Burstein, Chodorow, & Leacock, 2003) and are calculated as rates (total number of errors divided by number of words in the essay).
- *Organization and development:* e-rater identifies sentences in each essay corresponding to important parts of an essay (background, thesis, main ideas, supporting ideas, conclusion). The organization score is based on the number of discourse elements in the essay. The development variable is the average number of words of each of these elements in the essay.
- *Lexical complexity:* e-rater calculates two feature variables related to lexical characteristics—a measure of the vocabulary level of each word, based on a large corpus of newspapers and periodicals, and the average word length.
- *Prompt-specific vocabulary usage*: e-rater compares the vocabulary in each essay with the vocabulary used in essays at each of the score points on the rating scale and computes two variables. The first is the score point value, which calculates the score point to which the essay is most similar, and the second is the cosine correlation value, which indicates how similar the essay is to essays at the highest point on the rating scale.
- *Essay length*: In previous versions of e-rater, essay length was not explicitly included as a variable, but the current version includes essay length (number of words) so that its effect can be controlled, particularly in calculating error rates as described above.

Validity of Automated Scoring Systems

Like other commercially available automated scoring systems (e.g., see Elliot, 2003; Landauer et al., 2003; Page, 2003), e-rater has been demonstrated to be highly correlated with scores given by human raters (Burstein, 2002; Burstein & Chodorow, 1999). However, despite findings that automated scores are as reliable as human scores, use of automated scoring has generated controversy and strong opposition, particularly among composition teachers, primarily because a computer cannot actually read student writing (Anson, 2006; Herrington & Moran, 2001).

A recent position statement by the Conference on College Composition and Communication (CCCC, 2004) stated in part:

> The speed of machine-scoring is offset by a number of disadvantages. Writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes. If a student's first writing-experience at an institution is writing to a machine, for instance, this sends a message: writing at this institution is not valued as human communication—and this in turn reduces the validity of the assessment. Further, since we can not know the criteria by which the computer scores the writing, we can not know whether particular kinds of bias may have been built into the scoring. And finally, if high schools see themselves as preparing students for college writing, and if college writing becomes to any degree machine-scored, high schools will begin to prepare their students to write for machines. ("A Current Challenge" section, para. 2)

A number of scholars have expressed similar concerns about the consequences of automated scoring on the teaching and learning of writing. For example, Cheville (2004) noted that in the real world what counts as an error in one situation may be completely appropriate in another. The algorithms used in automated scoring have no way of taking into account the sociolinguistic context in which particular choices of vocabulary or syntax may be seen as errors or not, and they thereby give students the false idea that errors can be objectively defined and thus avoided.

Herrington and Moran (2001) argued further that relying on automated scoring systems as a replacement for on-campus placement programs will result in the loss of staff training that occurs on campus as faculty develop writing criteria: "So long as placement tests are developed in-house, there have to be conversations among faculty and administrators about what it means to be 'proficient'" (p. 496). Other concerns include the impact of automated scoring on the teaching and learning of writing (e.g., Cheville, 2004) and the constraints on the assessment task that are necessary for automated scoring to be feasible (e.g., Condon, 2006; see also Bennett & Bejar, 1998, for a more general discussion of task design considerations for automated scoring).

In terms of the TOEFL interpretive argument (Chapelle et al., 2008), the concerns raised by these scholars fall under the category of utilization: "The meaning of test scores is clearly interpretable by admissions officers, test takers, and teachers. The test will have a positive influence on how English is taught." (p. 21). Addressing these concerns directly is beyond the

scope of this particular study; however, inferences about utilization of test scores rely on a chain of evidence for intermediate inferences such as generalization and extrapolation, which are addressed in this study.

Yang, Buckendahl, Jusziewicz, and Bhola (2002) identified three main approaches to validating automated scores. One approach involves investigation of the relationship between automated scores and scores given by human raters. Another approach is to examine relationships between automated scores and external measures of the same ability (i.e., criterion-related validity evidence). The third approach is to investigate the scoring process and mental models represented by automated scoring systems (see for example Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; and Lee, Gentile, & Kantor, 2008 for examples of this line of research). The current study focuses on the first two of these approaches.

As noted previously, several studies have demonstrated the comparability of scores between human raters and automated scores (Yang et al.'s first category). One important study in this area is from Chodorow and Burstein (2004), who found that, once the effects of essay length were removed, e-rater v. 01 was not sensitive to certain characteristics of writing that human raters were. Chodorow and Burstein concluded that future improvements to e-rater should be made to capture some of these characteristics, including additional measures of syntactic proficiency and word usage. These measures have been included in the current version of e-rater, as noted above.

The literature in the second category, the criterion-related validity of automated scores, is scant, although some researchers have looked at the relationship between human scores on writing assessments and performance on other measures of writing. Breland, Bridgeman, and Fowles (1999) provided an overview of studies that have investigated the predictive validity of writing assessments ranging from in-house placement tests to the Law School Admissions Test (LSAT) and the SAT® Writing Subject Test. The criteria used for these studies have been (a) course grades, grade point averages, or instructors' ratings; (b) performance on other writing tasks (specifically, multiple essays scored by multiple raters); and (c) examinee background indices, including self-assessment of writing ability. Breland et al. found that essay test performance correlated more highly with other writing performance than with grades, GPA, or instructors' ratings.

Studies that have related automated scores to nontest indicators of writing ability include Elliot (as cited in Attali & Burstein, 2006) and studies by Peterson and by Landauer, Laham, Rehder, and Schreiner (both cited in Powers, Burstein, Chodorow, Fowles, & Kukich, 2000). A model for the current study was Powers et al. (2000). This study looked at correlations between e-rater and human scores on two essay tasks from the GRE® General Test with several other indicators of writing ability: (a) two samples of writing prepared as course assignments, (b) self-evaluations of writing, (c) self-reported grades in writing-intensive courses, (d) self-reported documentation of accomplishments in writing, and (e) success with various kinds of writing. The researchers found modest but significant correlations between e-rater scores and most of the indicators, with the highest correlations being with evaluators' grades on course assignments. E-rater did not fare as well as human raters in these correlations, one possible explanation being that the version of e-rater used in the study did not focus to the same degree as human raters on aspects of writing reflected in the nontest indicators and that e-rater tended not to assign extreme scores. This study suggests that the validity-related inference of generalizability across raters may not be fully supported for e-rater, at least in earlier versions.

Automated Scoring and Nonnative Writing

In addition to the issues raised above, there is another set of validity-related issues surrounding the use of automated scoring for nonnative writers. E-rater, like other automated scoring systems, was designed initially with a population of native speakers in mind. For it to be accepted as a valid method of scoring nonnative speakers (NNS) of English, particularly the population of TOEFL examinees, a number of considerations will need to be dealt with. One issue is computer familiarity of examinees—since automated scoring depends on digital rather than paper-and-pencil tests, evidence must be presented that the lack of keyboarding skills does not lead to construct-irrelevant variance. The issue of computer familiarity is of particular importance to the TOEFL because of variable access to computer technology in the different countries that comprise the population of TOEFL examinees. The question of computer familiarity as it relates to the TOEFL was first discussed in Kirsch, Jamieson, Taylor, and Eignor (1998), who found a relationship between level of computer familiarity and TOEFL scores.

Wolfe and Manalo (2004) found an interaction between language proficiency and chosen medium (handwriting or word processing), with lower proficiency students performing better if they handwrote their essays and higher proficiency students performing better if they input their

essays on the computer. Wolfe and Manalo expressed concerns that groups traditionally associated with low computer familiarity or higher computer-related anxiety (e.g., females, examinees from developing countries, and older examinees) tend to choose handwriting over word processing. If these examinees are required to use computers in writing assessments they may have to perform a "double translation," which increases the cognitive demands of the task—Not only do they have to translate from their native language into English, but they also have to translate from English into unfamiliar keystrokes. This additional cognitive load is a potential source of construct-irrelevant variance, and more research is needed to explore this issue.

Another set of concerns related to the assessment of NNS writing is the question of whether the features used to score essays, particularly in the areas of grammar, usage, and vocabulary, are in fact the features of language that are problematic for NNS. Since the Criterion analysis tools used to detect errors in grammar and usage are intended to focus on the kinds of errors typically made by native speakers rather than those found in NNS texts (Burstein et al., 2003), the errors extracted by Criterion and thus used in e-rater are not necessarily those that appear in NNS writing. However, it should be noted that work is being done to improve identification of typical NNS errors such as prepositions and articles (Chodorow, Gamon, & Tetreault, 2010).

Another issue to be taken into consideration is the fact that the TOEFL differs from other writing tests used for screening and university admission in that it is a test of language proficiency rather than an aptitude test or a test of analytical thinking. Indeed, research on second-language writing (e.g., Cumming, 1989; Sasaki & Hirose, 1996) suggests that language proficiency and writing ability are separate, although related, constructs. While predictive validity studies of tests such as the SAT, GMAT, and GRE Tests presume that the ability measured by the test is more or less stable, this is not the case for the TOEFL. As Simner (1999) noted:

> The major purpose of using the TOEFL as an admissions screening device is not to determine how well a student performs in English at the time the TOEFL is taken, but instead to determine how well the student is likely to perform in the future, which typically means some 8-10 months later after the student has arrived on campus and is immersed in an English speaking environment. Hence, the evidence needed to support the TOEFL as a screening device is evidence in favor of predictive validity. (p. 287)

Studies of the predictive validity of the TOEFL have had mixed results. A few studies have looked at the relationship between TOEFL scores and indicators of success such as graduation rate, GPA, or GPA after the first 9 credit hours. For example, Ayres and Peters (1977) found that TOEFL scores were predictive of graduate grade point average (GGPA) among Asian students in science and engineering, and that a combination of TOEFL and the verbal section of the GRE General Test predicted success in program completion. On the other hand, Neal (1998) found no relationship between TOEFL scores and GGPA. Studies by Light, Xu, and Mossop (1987), Xu (1991), and Yule and Hoffman (1990) also found little evidence of a relationship between TOEFL scores and academic success. It should be noted that these studies were based on the total TOEFL score, not the writing score in particular; little attention has been paid to the predictive validity of the TOEFL writing test specifically.

One reason that the TOEFL in general does not consistently demonstrate predictive validity is that language proficiency in itself is only one of many factors that influence success in university studies. Another reason is that requirements for language skills and proficiency may vary by college and major, so that students with lower TOEFL scores may be successful in some areas and not in others. A third reason is that different levels of support for international students with limited proficiency are offered at different institutions. For these reasons it is not likely that TOEFL scores by themselves will ever be strongly predictive of academic success, beyond providing a threshold (floor) below which students have a strong probability of not being successful because of limitations in their language proficiency.

To summarize, using automated scoring systems for the TOEFL, which is intended for nonnative writers, brings up certain validity questions beyond those that may apply to tests of writing for native speakers. The research described here does not attempt to answer all of these questions; however, these questions should be kept in mind when interpreting research results and planning additional research in this area.

Task Variability in Writing Assessment

The advantages of a direct test of writing, as opposed to a more indirect test such as a multiple-choice test, come with the serious disadvantage of a limited ability to sample the domain adequately, so that writing tests are often limited to a single 30-minute prompt. It is therefore critical to ensure that differences across prompts are minimized so that examinees have an equal chance of performing successfully on all potential tasks. Task variability can affect performance

in a number of ways (see Weigle, 2002, chap. 4, for an overview); in the words of Purves (1992): "different tasks present different problems, which are treated differently by students and judged differently by raters" (p. 112).

Because each TOEFL examinee writes on only a single independent topic, there has been little opportunity to investigate the reliability of scoring (human or automated) across different topics using data from the same people. Most studies of writing prompts from the TOEFL, or its predecessor, the Test of Written English™ (TWE®), rely on other means of analyzing prompt-related differences. For example, an early study done of essays written for the TWE found small but significant differences across eight different prompts, studying the operational administration of these prompts worldwide (Golub-Smith, Reese, & Steinhaus, 1993). In more recent studies applying e-rater to TOEFL essays, neither Burstein and Chodorow (1999) nor Chodorow and Burstein (2004) used essays written by the same people in their studies of applying e-rater to nonnative speakers of English. Attali (2007, 2008) is a notable exception, in that he investigated the reliability of human and e-rater scores of essays for repeat test takers; however, Attali's study did not look specifically at task-related differences.

Despite not using essays written by the same candidates, Chodorow and Burstein (2004) found that scores of human raters were more variable across prompts than were automated scores, and also found a significant main effect of prompt on essay scores, significant main effects of rater (human vs. two versions of e-rater) and native language, and interactions between prompt and rater, rater and language, and language and prompt. It appears, therefore, that investigating effects of differences among TOEFL prompts is still an area where more research is needed.

The standardized writing features included in e-rater offer an opportunity to investigate differences in the textual structure of essays written to different prompts by the same candidates. Attali and Burstein (2006) used essays from the Criterion database written by students from 6th through 12th grades to investigate reliability across essay prompts, and found that e-rater and human scores were very highly correlated. Furthermore, they found that certain features used by e-rater had moderate test-retest reliabilities, most of which were in the mid .40s. No study to date has looked at differences in e-rater feature scores across prompts of the TOEFL.

To summarize the literature review, I will return to the TOEFL interpretive argument articulated by Chapelle et al. (2008). In terms of generalizability, the literature suggests that improvements in e-rater have reduced the gap between automated scores and human scores

considerably, though some questions remain about this equivalence for the TOEFL in particular. Furthermore, there is little research comparing performance by the same students on different TOEFL writing tasks both on overall scores (human and e-rater) and e-rater features. In terms of extrapolation, there is a dearth of research addressing the relationship between the construct of writing assessed by the TOEFL (and embodied in the scoring rubric used by human raters and the algorithms used by e-rater) and the actual writing performance of students outside the testing construct. The study reported on here attempts to address these issues.

Method

Research Questions

This study addresses the following research questions:

1. What are the relationships between overall e-rater and human scores on TOEFL iBT independent writing tasks and other indicators of writing ability (self-assessment of writing ability, instructor assessment of writing ability, and independent rater assessment on discipline-specific writing tasks)?
2. What are the relationships among specific features analyzed by e-rater and these indicators of writing ability?
3. How consistent are the scores generated by e-rater (both the total scores and scores for individual features) and human raters across two different writing tasks?

Participants

Data were gathered from 386 nonnative English-speaking students at eight different institutions in the US over a 15-month period, from October 2006 through December 2007 (see Table 1 for participant characteristics). Participants were recruited from the international student populations at the following institutions: Iowa State University, Georgia State University, Michigan State University, Pace University, the University of California at Los Angeles, Purdue University, Portland State University, and the University of Minnesota. The original intention was to test matriculated students only, but at one institution 26 students enrolled in that university's English Language Institute were included in the participant pool. Participants were each paid $50 for their participation, in the form of a gift card for their university bookstore.

Table 1

Participant Characteristics

	Characteristic	*N*
Total		386
Age	Mean (years): 24.86 Range: 18–47	
Gender	Female	222
	Male	163
	Unknown	1
Native language	Chinese	158
	Korean	51
	Japanese	25
	Spanish	17
	Vietnamese	13
	Russian	13
	French	11
	Turkish	10
	Other	88
Status	Graduate	199
	Undergraduate	159
	Other (ELI[a]/not specified)	28
Field of study	Business	93
	Social Sciences	88
	Engineering	49
	Humanities	41
	Natural Sciences	37
	Computer Science	22
	Education	15
	Applied Sciences	12
	Health Sciences	12
	Mathematics	8
	Missing/Other	9

[a] English language institute.

Materials

The following data were collected:

Essays responding to TOEFL iBT tasks. Two independent writing tasks, provided by ETS for this study, were administered to participants. One prompt (hereafter referred to as Topic 1) asked students to discuss whether too much emphasis is spent on personal appearance and fashion, and the other (hereafter referred to as Topic 2) dealt with the importance of planning for the future. The order of prompts was counterbalanced so that half of the participants received one prompt first and half received the other prompt first.

Self-assessment of writing ability. A web-based survey adapted from Allwright and Banerjee (1997) was created using SurveyMonkey, an online survey development tool (see Appendix A for the survey). The student survey had four sections. First, students were asked to rate their ability to write, read, speak, and understand English and also to compare their ability to use English for coursework with their ability to use English outside of school. Next, students were given a list of nine problems that students sometimes have with writing and were asked to indicate how often they experienced these problems. In the third and fourth sections, respectively, students were asked to indicate how often they experienced specific problems related to other aspects of English (e.g., speaking, reading, and participating in class discussions) and to nonlanguage related problems (e.g., time management and understanding the subject matter). In each section students could provide open-ended comments as well.

To validate the student survey, a factor analysis of the survey data (excluding the overall self-evaluation variables) was conducted using principal components analysis with varimax rotation (see Appendix B). The factor analysis revealed three main factors similar to the intended factors, with the exception of three writing items that loaded on the third factor instead of the writing factor. Accordingly, the following three scales were constructed: (a) Writing problems (6 items, $\alpha = .82$), (b) Other language problems (5 items, $\alpha = .81$), and (c) Other problems (7 items, $\alpha = .80$).

Instructor assessment of writing ability. Participants were asked to provide names and contact information for two instructors familiar with their written work. These instructors were contacted by e-mail and asked to complete an online survey (see Appendix C for the survey). The instructor survey was similar in structure to the student survey, with sections asking instructors to

rate the student's overall performance in the course, the student's writing ability, oral ability, and overall level of English, and their perceptions of the impact of linguistic and nonlinguistic factors on the student's performance. Instructors also were invited to make open-ended comments in each section of the survey.

As with the student survey, a factor analysis was conducted to explore the structure of the survey (see Appendix D). Two scales were constructed, one for language-related problems (9 items, $\alpha = .96$) and one for nonlanguage related problems (5 items, $\alpha = .91$). Unlike the student survey, which had very few missing responses, many instructors chose the option "no opportunity to judge" on several items, which was recorded as a missing response. Therefore each scale score was calculated as the average of nonmissing scores rather as the total of the nonmissing scores.

Nontest writing samples. Participants were asked to provide two samples of writing for courses in which they had been enrolled within the past 6 months. Participants were encouraged to provide writing samples from their major courses, but many only had writing samples from writing classes (i.e., English composition or English as a Second Language [ESL] courses). Participants were asked to provide, if possible, one sample that represented their typical writing and one that was not as good as their typical writing. The rationale for this request was based on Powers et al.'s (2000) observation that students tend to submit their best samples, rather than typical samples; thus an attempt was made to obtain writing that was more representative of typical course-related writing. Approximately half the collected samples were from major courses and half were from English composition or ESL courses. Samples of student writing are found in Appendix E.

Participant information sheet. The participant information sheet (see Appendix F) served two functions. First, it provided an opportunity to collect basic demographic information from students. Second, it served as the vehicle for collecting contact information for students' instructors and information about the two writing samples students were asked to provide. This information included the name of the course for which the paper was written, their estimation of the strength of the writing, and the types of assistance, if any, students had received on the paper.

Procedures

When participants signed up for the study they were given the information sheet and asked to bring it back completed on the test date, along with their two writing samples. When they arrived at the testing site, they logged on to a secure website, where they took the student survey and then the writing test. The two writing topics were presented in random order. When the students had completed all study requirements, including supplying contact information for their instructors and submitting their writing samples, they were compensated and then dismissed.

Following student data collection, instructors were contacted by email with a request to complete the instructor survey. Reminders were sent to instructors after 2 weeks; in some cases a second reminder email was sent to instructors who had not yet completed the survey. A total of 410 instructors completed the survey; of the 386 student participants, 186 (48%) had one instructor response, 112 (29%) had two, and 88 (23%) had no instructor information.

Scoring of iBT essays. All TOEFL iBT essays were sent to ETS for scoring with the current version of e-rater. The generic or "program specific" e-rater model uses eight features and was built on the responses of tens of thousands of examinees to more than 25 TOEFL prompts, including the two prompts used in this study (Attali, 2007). The only prompt-specific customization of the model was that the machine scores were scaled to have the same mean and standard deviation as human ratings for the specific prompt. The e-rater features used in the study were the features described above, except that the two prompt-specific vocabulary scores and essay length were not included.

Each TOEFL iBT essay was also scored by two trained raters using the TOEFL scoring rubric (see Appendix G). The TOEFL iBT essays were also scored by trained raters. Approximately half of the scripts were scored by experienced raters certified by ETS; a total of four raters participated in the first round of rating. The second half of the scripts were rated by raters hired by the author; they were experienced ESL teachers who had rated other writing assessments but not TOEFL essays. These raters completed the ETS online training before rating the scripts but were not certified by ETS. The author also rated any essays that received scores from the two human raters that were more than one point apart; however, all analyses presented in this report are based on the scores of the original two raters. For all analyses involving individual raters, ratings have been randomly assigned to Rater 1 or Rater 2. Table 2 shows interrater reliability statistics for these ratings; overall, they are comparable to statistics found in similar

studies (e.g., Attali, 2007; Attali & Burstein, 2006). For example, Attali and Burstein (2006) reported exact agreement rates of two human raters of .59 and one human rater with e-rater of .58. Pearson correlations between individual raters (i.e., not ratings) ranged from .54 to .83; correlations of individual raters with e-rater scores ranged from .66 to .75. Across the two topics, correlations were as follows: Rater 1, $r = .62$; Rater 2, $r = .58$; Average rating, $r = .71$; e-rater, $r = .79$. This suggests that e-rater is somewhat more reliable than human ratings in terms of alternate-forms reliability.

Table 2

Interrater Reliability of e-rater and Human Rater Scores

	Topic 1	Topic 2	Overall
Rater 1/Rater 2			
Pearson correlation	.67	.64	.65
Exact agreement/exact + adj. agreement	.57/.97	.47/.94	.52/.96
Rater 1/e-rater			
Pearson correlation	.67	.75	.71
Exact agreement/exact + adj. agreement	.52/.97	.51/.98	.52/.98
Rater 2/e-rater			
Pearson correlation	.71	.72	.72
Exact agreement[a]/exact + adj. agreement	.56/.96	.49/.97	.53/.97
Average of 2 HR/ e-rater			
Pearson correlation	.76	.81	.79
Exact agreement[b]/exact + adj. agreement	.73/.95	.76/.97	.74/.96

Note. Exact agreement means that the two raters gave exactly the same score; adjacent agreement means that the two scores differed by one point or less. For analyses involving e-rater, scores were rounded off to the nearest whole number. Since the average of two human rater scores was not always a whole number, agreement was counted as exact if the rounded e-rater score was within ½ point of the average of two raters.

Scoring of submitted writing samples. The course writing samples provided by students were scored by a pool of trained raters on a scale designed for the study consisting of two subscales: content and language (see Table 3). Each sample was rated by two raters, with a third rater adjudicating if the two raters differed by more than a point on either scale. The reported score is the average between the two raters. Pearson correlations between the two (averaged) ratings on each scale across samples were .51 for content and .58 for language; within-samples correlations between content and language were .78 for Sample 1 and .79 for Sample 2. The original scale included two score points below *Fair* but as no submitted samples were judged to be below *Fair* these two points were excluded from the final rating scale.

Table 3

Scoring Rubric for Submitted Writing Samples

Score	Content	Language
6 – Excellent	Issues dealt with fully, clear position, substantive arguments, balanced ideas with full support and logical connection, strong control of organization	Excellent control of language with effective choice of words, sophisticated range of grammatical structures and vocabulary, few or no errors
5 – Very good	Issues dealt with well, clear position, substantive arguments, generally balanced ideas with support and logical connection, good control of organization, occasional repetition, redundancy, or a missing transition	Strong control of language, read smoothly, sufficient range of grammatical structures and vocabulary with occasional minor errors
4 – Good	Issues discussed but could be better developed, positions could be clearer and supported with more substantive arguments, appropriate organization, with instances of redundancy, repetition, and inconsistency	Good control of language with adequate range of grammatical structures and vocabulary, may lack fluidity, some grammatical errors
3 – Fair	Issues discussed, but without substantive evidence, positions could be clearer and arguments could be more convincing, adequate organization, ideas are not always balanced	Fair control of language with major errors and limited choice of structures & vocabulary, errors may interfere with comprehension

Data analysis. For Research Question 1 the relationship between essay scores and criterion variables was investigated primarily through correlations. Pearson correlations between criterion variables (student survey variables, instructor survey variables, and ratings on nontest writing samples) and ratings on TOEFL iBT essays were computed separately for each prompt as follows:

- E-rater (ER); 1 rating per prompt (2 total)
- Each human rating (1 HR); 2 ratings per prompt (4 total)
- The average between the two raters (2 HR); 1 averaged rating per prompt (2 total)
- The average of each human rating and e-rater (1 HR/ER); 2 averaged ratings per prompt (4 total)

The average of the correlations in each category across rater combinations and prompts (single human rater, average of two human raters, e-rater, and average of one human rating and e-rater) is reported in the results.

Where appropriate, differences in the magnitude of correlations between e-rater scores and the average of two human rater scores, respectively, with criterion variables were calculated using procedures outlined in Cohen and Cohen (1983, p. 57; see Urry, 2003, for the SPSS syntax).[1] Operationally, e-rater is used as one of the two raters for the TOEFL; however, it was designed to emulate the average of two raters' scores. For this reason the average between the two raters was felt to be the most appropriate human rating to compare with e-rater for this analysis.

For Research Question 2 the e-rater feature scores were averaged across the two writing prompts, and Pearson correlations were calculated among the averaged features scores and criterion variables (global self and instructor ratings of language ability and scores on nontest writing samples). Finally, for Research Question 3 paired *t*-tests were conducted to compare scores on the two prompts in terms of individual ratings, the average human rater scores, e-rater total scores, and feature scores by prompt. In addition, a repeated-measures ANOVA was conducted with rater and prompt as independent variables and score as the dependent variable. All statistical analyses were carried out using SPSS Versions 15 and 16.

Research Question 1: Results and Discussion

Research Question 1 (regarding the relationship between human and e-rater scores and other indicators of writing ability) was addressed through correlations between scores on iBT essays (human and e-rater) and a variety of criterion variables. As noted above there were three main data sources apart from the TOEFL essays: student surveys, instructor surveys, and ratings on other writing samples. For the student and instructor surveys, correlations were calculated between TOEFL essay scores and both the global evaluation items and the survey scales as described above. For the additional writing samples, correlations were calculated between TOEFL essay scores and scores on content and language.

Results

Relationships between essay scores and student self-assessment. The relationships between scores on TOEFL iBT essays and student survey variables are presented in two sections: First student overall self-evaluations of language ability are discussed, and then specific problems that are related to language as well as those that are not. In the survey, students were asked to rate their ability in the skills of writing, speaking, listening, and reading on a scale of 1 to 4. Descriptive statistics for these variables are found in Table 4, and correlations among these variables and the problem scales from the survey are found in Table 5. As Table 4 shows, students rated themselves the highest in receptive skills (reading and listening) and lowest in productive skills (writing and speaking). Table 5 shows that the global self-evaluation variables have moderately strong correlations with each other (.59 to .69) but are less strongly related to the three problem scales (.33 to .49); the correlations among the scale variables themselves range from .52 to .60.

Average correlations between combinations of e-rater and human scores on the TOEFL iBT essays and self-evaluation variables are found in Table 6. As noted above, these correlations are averaged across the two prompts for e-rater and the average human rater score and across both raters and prompts for single human rater scores. The correlations are moderate, with higher correlations for reading and writing than for listening and speaking.

Table 4

Descriptive Statistics of Global Self-Evaluation Variables

	N	Mean	SD
Self-Evaluation Writing	382	2.63	0.83
Self-Evaluation Reading	382	2.96	0.84
Self-Evaluation Listening	378	3.05	0.84
Self-Evaluation Speaking	381	2.66	0.85

Table 5

Correlations Among Global Self-Evaluation Variables and Student Survey Scales

	Global Self-Evaluation variable	1	2	3	4	5	6	7
1.	Self-Evaluation Writing	—	.68	.59	.68	.48	.41	.40
2.	Self-Evaluation Reading		—	.69	.60	.43	.49	.42
3.	Self-Evaluation Listening			—	.65	.37	.49	.33
4.	Self-Evaluation Speaking				—	.41	.49	.33
5.	Writing Problems Scale					—	.52	.57
6.	Language Problems Scale						—	.60
7.	Other Problems Scale							—

Table 6

Average of Correlations Between Scores on TOEFL iBT Tasks and Self-Evaluation Variables

	e-rater	1 HR	2 HR	1 HR/ER
Self-Evaluation Writing	.36	.39	.43	.41
Self-Evaluation Reading	.36	.38	.42	.40
Self-Evaluation Listening	.23	.31	.33	.29
Self-Evaluation Speaking	.26	.32	.35	.31

Note. All individual correlations were significant at $p < .01$. 1 HR = individual human rating; 2 HR = average between the two raters; 1 HR/ER = average of each human rating and e-rater.

Table 7 displays the results of a *t*-test comparing the differences in the magnitude of correlations between e-rater scores and the average of two human rater scores, respectively, with the self-evaluation variables as described above. As the table shows, the correlations with the human rater scores were significantly higher than those with most of the corresponding e-rater scores, although the effect sizes ($r^2_1 - r^2_2$, Cohen, 1998, pp. 114-115) are small.

Descriptive statistics and correlations with ratings for the three scales are found in Table 8. Students reported the most problems with writing (mean = 17.98 out of 24, or 75% of the maximum) and the least with other (nonlanguage-related) problems (Mean = 23.67 out of 28, or 85% of the maximum); note that a higher mean score represents fewer problems than a lower mean score. The table also shows that human and e-rater scores were moderately and similarly related to the student survey variables and that the correlations were lower than the correlations with overall self-evaluation variables discussed above.

Relationship between scores and instructor assessment of writing ability. As noted earlier, 296 of the 386 student participants received at least one instructor survey assessment. For the purposes of this analysis only the responses for the first instructor who responded for each student have been analyzed; however, because 112 students had two instructor responses it is possible to look briefly at how the two instructors' responses compared with each other. Pearson correlations between the two instructors' ratings on individual survey items and scale scores were quite low, in some cases close to 0. The low correlations may be explained partly by the fact that most ratings were at the high end of the scale, resulting in a restricted range. A more accurate measure of the interrater reliability is thus the percentage of exact and adjacent agreement; in other words, how often did the two instructors agree (or come close to agreeing) on their ratings of individual students? Cross-tabulations of the ratings reveal that exact agreement varied from 45% to 50% and that exact-plus-adjacent agreement ranged from 80% to 95%, thus indicating acceptable interrater reliability by this measure.

Another important factor to consider when interpreting the low correlations between the two instructors is the content area of the instructors. The correlations were generally much higher when both instructors were either English/ESL teachers or content teachers and lower (even negative) when one instructor was an English/ESL teacher and the other was a content teacher. For example, correlations on the Language Problem scale were .45 ($p < .01$) when both

Table 7

t-test of Difference in Magnitude of Correlations of e-rater and the Average of Two Human Raters With Self-Evaluation Variables

	Topic 1			Topic 2		
	N	*t (p)*	Effect size[a]	*N*	*t (p)*	Effect size[a]
Self-Evaluation Writing	367	0.24 (.81)	.01	370	-3.55 (.00)	.07
Self-Evaluation Reading	367	-1.62 (.05)	.04	370	-2.23 (.01)	.05
Self-Evaluation Listening	367	-3.50** (.00)	.07	370	-3.17 (.00)	.05
Self-Evaluation Speaking	367	-2.29* (.01)	.04	370	-3.40 (.00)	.07

[a] Effect size is calculated as $r^2_1 - r^2_2$ following procedures outlined in Cohen (1988, pp.114–115). Effect sizes lower than .09 are considered small.

Table 8

Descriptive Statistics and Average Correlations Between Scores on TOEFL iBT Tasks and Student Survey Scales (Higher Mean Scores = Fewer Problems)

	Descriptive statistics					Average correlations[a]			
	N	Range	Mean	SD	Reliability[b]	e-rater	1 HR	2 HR	1 HR/ER
Writing problems	367	8–24	17.98	3.37	.82	.30	.27	.29	.31
Other language problems	381	3–20	16.03	3.09	.81	.14	.17	.19	.17
Other problems	344	9–28	23.67	3.36	.80	.25	.23	.26	.26

Note. All individual correlations significant at $p < .01$ except between e-rater and other language problems, which was significant at $p < .05$ on Topic 1 and not significant on Topic 2. 1 HR = individual human rating; 2 HR = average between the two raters; 1 HR/ER = average of each human rating and e-rater.

[a] Correlations of e-rater and average human rater scores with criterion variables were not significantly different. [b] Cronbach's alpha.

instructors were English/ESL teachers, .12 (*ns*) when neither instructor was an English/ESL teacher, and -.23 (*ns*) when the two were an English/ESL instructor and a content instructor This suggests that the language demands of different English/ESL courses may be more similar to each other than they are to those of content area courses or than those of content courses are to each other. It also explains the near-zero correlations when instructors are not grouped in this way.

Of the 296 instructors who were the first respondents to the survey, slightly more than 50% (153) were English, writing, or ESL instructors, and the rest (143) were subject instructors. The responses for these two groups were analyzed separately for a number of reasons. As noted above, perhaps these two groups appeared to respond differently to the survey items because writing courses in general, and ESL writing courses in particular, focus on the mastery of language-related skills rather than specific knowledge about an academic discipline. In the courses, assignments are adjusted with respect to the presumed writing and/or language ability of students in the class. In a lower-level ESL course, for example, the readings and writing assignments may be shorter and simpler than for a higher level writing course or a course in an academic discipline such as philosophy or business. Thus the responses to an item that asks instructors to judge, for example, whether a student has problems understanding course assignments will likely be different between these two groups of instructors.

For the purposes of this study, perhaps the most important reason to distinguish between these two instructor groups is that the readings, assignments, and other demands of content course areas represent, in fact, the target language use situation (Bachman & Palmer, 1996) of the TOEFL. That is, test users (e.g., admissions officers) are interested in knowing how well prospective students will be able to use English in academic disciplines such as biology, economics, or psychology. Thus in the investigation of the predictive validity of the TOEFL it is particularly important to distinguish between instructors of content courses and English or ESL instructors when examining instructor perceptions of NNS performance in their courses.

Like the student survey, the instructor survey included both global assessments of language proficiency and items asking about specific problems that students may face. Descriptive statistics for the proficiency variables are found in Table 9, and intercorrelations between these variables and the instructor survey scale variables are found in Table 10.

Table 9

Descriptive Statistics for Instructor Ratings of Overall Performance and Proficiency

	Subject			English			Overall		
	N	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Overall academic performance	138	3.23	0.63	153	2.99	0.79	291	3.11	0.73
Writing ability evaluation	142	3.06	0.76	146	2.82	0.87	288	2.94	0.82
Oral proficiency evaluation	143	3.16	0.80	153	2.92	0.86	296	3.04	0.84
General evaluation of English ability	143	3.14	0.70	151	2.89	0.80	294	3.01	0.76

Table 10

Correlations Among Instructor Survey Variables

	1	2	3	4	5
1. Writing ability	—	.61	.83	.44	.40
2. Oral proficiency		—	.79	.40	.46
3. Overall ability			—	.42	.48
4. Language problems				—	.43
5. Other problems					—

A few observations can be made from these tables. First, Table 9 shows that ratings by subject area instructors were higher than ratings by English instructors; this is not surprising, since many students in the study were specifically placed into English/ESL classes because of a perceived need to improve their English. Table 10 shows that the instructor evaluations of different aspects of English proficiency were more highly correlated than the similar self-evaluation variables discussed earlier; however, the relationship between the overall evaluations of proficiency and the scale variables were approximately the same for both the student and instructor surveys.

Correlations between the overall instructor evaluation variables and the TOEFL iBT ratings are found in Table 11. Interestingly, scores on TOEFL iBT essays correlated more strongly with subject area instructor ratings of student proficiency than with those of English instructors; this result may be due to the differences in comparison groups noted above. E-rater correlations were slightly lower than human rater correlations, but these differences were significant only for Topic 1 for the overall evaluation (t = -2.91, df = 285, p < .01, effect size = .08).

Table 11

Average of Correlations Between Scores on TOEFL iBT Tasks and Instructor Ratings of Overall Performance and Proficiency

	e-rater	1 HR	2 HR	1 HR/ER
Subject (n = 138–143)				
Overall academic performance	.23	.24	.25	.26
Writing ability	.30	.34	.37	.35
Oral proficiency	.36	.39	.41	.41
Overall English ability	.38	.42	.46	.44
English (n = 146–153)				
Overall academic performance	.13	.13	.14	.14
Writing ability	.22	.20	.21	.22
Oral proficiency	.16	.19	.20	.18
Overall English ability	.27	.27	.29	.29
Total (n = 288–296)				
Overall academic performance	.21	.22	.23	.23
Writing ability	.28	.29	.32	.31
Oral proficiency	.27	.31	.33	.31
Overall English ability	.34	.36	.39	.38

Note. n refers to the sample sizes for individual correlations, which vary within each category because of missing data. For sample sizes 138–153, individual correlations below approximately .16 are not significant, between .17 and .21 significant at $p < .05$, and above .21 at $p < .01$. For samples sizes 288–296, individual correlations above .16 are all significant at $p < .05$. 1 HR = individual human rating; 2 HR = average between the two raters; 1 HR/ER = average of each human rating and e-rater.

As noted earlier, the instructor survey consisted of nine language-related questions and five nonlanguage-related questions. Descriptive statistics for these scales are found in Table 12, and correlations with TOEFL iBT essay ratings are presented in Table 13.

As was the case with the overall proficiency and performance variables, the scale scores from subject area instructors were generally higher than those from English instructors. In addition the correlation between the language impact scale and TOEFL essay scores was significantly higher for the average of two human raters than for e-rater for the subject area

Table 12

Descriptive Statistics for Instructor Survey Scale Variables

	N	Range	Mean	*SD*
Total				
Language impact	296	1–5	3.56	0.98
Impact of other factors	279	1–5	4.11	0.76
Subject				
Language impact	143	1–5	3.72	0.95
Impact of other factors	134	1–5	4.18	0.73
English				
Language impact	153	1–5	3.41	0.98
Impact of other factors	145	1–5	4.05	0.80

Table 13

Average Correlations of Instructor Survey Scale Variables With e-rater and Human Ratings of TOEFL iBT Essays

	e-rater	1 HR	2 HR	1 HR/ER
Subject (n = 133–143)				
Language impact[a]	.15	.31	.33	.26
Impact of other factors	.16	.21	.23	.20
English (n = 145–153)				
Language impact	.18	.15	.17	.18
Impact of other factors	.00	.01	.01	.00
Total (n = 279–296)				
Language impact	.21	.26	.28	.25
Impact of other factors	.09	.12	.13	.11

Note. n refers to the sample sizes for individual correlations, which vary within each category because of missing data. For sample sizes 133–153, individual correlations below approximately .16 are not significant, between .17 and .21 significant at $p < .05$, and above .21 at $p < .01$. For samples sizes 279–296, individual correlations above .16 are all significant at $p < .05$. 1 HR = individual human rating; 2 HR = average between the two raters; 1 HR/ER = average of each human rating and e-rater.

[a] Correlations of e-rater and average human rater scores with this variable were significantly different for both topics; no other correlations were significantly different.

instructors (Topic 1: t = -2.32, df = 138, p < .05, effect size = .09; Topic 2: t = -3.15, df = 138, p < .01, effect size = .07) but not for the English instructors.

Relationship between essay scores and scores on student-supplied writing samples. The third main indicator of writing ability examined in this study were the ratings on content and language for course-related writing samples provided by student participants. Descriptive statistics and interrater agreement statistics between the first two (unadjudicated) scores for these variables are found in Table 14. Recall that a third rater was used in those cases where the scores diverged by more than a point; the data in Table 14 do not include any of these third ratings. The reported scores in the table are the average scores between two raters.

As noted above, approximately half of the writing samples were from English or writing courses (e.g., ESL writing courses, technical writing) and half were from subject-area courses (e.g., chemistry, anthropology, applied linguistics). Recall that subject-area writing samples tended to be higher in both register and cognitive demands than English writing samples. Scores on subject-area papers were slightly higher than those on English/writing papers for both content (t = -8.61, df = 367, p = .000) and language (t = -8.58, df = 367, p = .000).[2] Correlations between scores on writing samples with e-rater, single human rater scores, and two human rater scores are found in Table 15. The table presents correlations for all samples combined as well as for samples divided into subject area versus English. There were no significant differences in the magnitude of correlations between the criterion variables and e-rater versus the average human score.

Note that the correlations in Table 15 are averaged across writing samples (Sample 1 and Sample 2) and prompts (Topic 1 and Topic 2). As the table shows, both e-rater and human rater scores were more highly correlated with scores on English papers than on subject area papers, and were more highly correlated with the language scores than the content scores. Overall the correlations tended to be higher between scores on nontest writing samples and TOEFL independent writing tasks than for other indicators of writing ability; furthermore, the correlations with e-rater scores appear to be more similar to those with human scores on this measure than with the other indicators.

Summary of results for Research Question 1. As a summary of the highest correlations between e-rater and essay scores and criterion variables, Table 16 displays the average correlations between e-rater, the average human rater score, and all variables where at least one correlation is greater than or equal to .3, sorted in descending order of the average human rater

score within each data source. Across most variables in the three data sources, the average correlations for e-rater and individual human raters were very similar, although the differences between e-rater and the average of two human rater scores were significant on the self-evaluation variables and some of the instructor survey variables. The most robust difference between e-rater and human ratings was in the language impact scale for subject area instructors.

Table 14

Descriptive Statistics and Interrater Agreement Statistics for Submitted Writing Sample Scores

	Descriptive statistics				Interrater agreement indices			
	N	Range	Mean	*SD*	*r*	Exact	Exact + adj.	Kappa
Content	748	3–6	5.13	0.70	.70	69.6	99.1	.51
Language	748	3–6	4.89	0.72	.71	64.7	99.9	.46

Note. Exact agreement means that the two raters gave exactly the same score; adjacent agreement means that the two scores differed by one point.

Table 15

Average Correlations of Submitted Writing Sample Scores With TOEFL iBT Task Scores by Content Area of Writing Samples

	e-rater	1 HR	2 HR	1 HR/ER
Content				
English	.39	.34	.37	.39
Subject	.23	.21	.24	.23
Total	.38	.36	.40	.40
Language				
English	.41	.38	.42	.43
Subject	.29	.30	.33	.32
Total	.42	.42	.46	.45

Note. All individual correlations significant at $p < .01$. 1 HR = individual human rating; 2 HR = average between the two raters; 1 HR/ER = average of each human rating and e-rater.

Table 16

Summary of Highest Average Correlations From Different Data Sources

Source	Variable	ER	1 HR	2 HR	1 HR/ER
SS	Self-Evaluation Writing[a]	**.36**	**.39**	**.43**	**.41**
SS	Self-Evaluation Reading[a]	**.36**	**.38**	**.42**	**.40**
SS	Self-Evaluation Speaking[b]	.26	**.32**	**.35**	**.31**
SS	Self-Evaluation Listening[b]	.23	**.31**	**.33**	.29
SS	Writing Problems Scale	**.30**	.27	.29	**.31**
IS	General Evaluation of English Ability (subject)	**.38**	**.42**	**.46**	**.44**
IS	Oral Proficiency Evaluation (subject)	**.36**	**.39**	**.41**	**.41**
IS	General Evaluation of English Ability (all)[a]	**.34**	**.36**	**.39**	**.38**
IS	Writing Ability Evaluation (subject)	**.30**	**.34**	**.37**	**.35**
IS	Language Impact Scale (subject)[b]	.15	**.31**	**.33**	.26
IS	Oral Proficiency Evaluation (all)[a]	.27	**.31**	**.33**	**.31**
WS	Content (all essays)	**.38**	**.36**	**.40**	**.40**
WS	Content (English essays only)	**.39**	**.34**	**.37**	**.39**
WS	Language (subject essays only)	.29	**.30**	**.33**	**.32**

Note. Correlations above .30 are in boldface. SS = student survey; IS = instructor survey; WS = writing samples; 1 HR = individual human rating; 2 HR = average between the two raters; 1 HR/ER = average of each human rating and e-rater.

[a] Correlation between criterion variable and 2 HR average was significantly higher than corresponding correlation with e-rater on one topic. [b] Correlations between criterion variable and 2 HR average were significantly higher than corresponding correlations with e-rater on both topics.

Discussion

Relationship between e-rater and human scores on TOELF iBT essays. From the results of the study it is clear that e-rater scores and human scores are highly correlated and thus can be said to be measuring highly similar constructs. From a practical perspective there seems to be little or no difference in scores between human raters and e-rater, and in fact the alternate-forms reliability of e-rater is somewhat superior to that of human raters in this study. However, there were some differences between e-rater and human scores on a few of the variables. These variables tended to be related to overall language proficiency rather than writing per se. For example, although the correlations between essay scores and self-evaluations of reading and writing were higher than those between essay scores and self-evaluations of listening and speaking, the correlations between human ratings and the latter self-evaluation scores were significantly higher than the corresponding correlations with e-rater scores. This finding suggests the possibility that the e-rater algorithm may not be as sensitive as human raters to certain markers of language proficiency.

The most striking difference between e-rater scores and the corresponding human scores is found in the relationship with subject instructors' ratings of the problems that their NNS students have that are related to language proficiency. One possible explanation for this result may be found in the research finding that essay raters do not base their scores strictly on the wording of a specific scale (see Eckes, 2008, for a recent review of the literature on rater behavior). For example, Lumley (2002) noted that raters' judgments seem to be based on "some complex and indefinable feeling about the text, rather than the scale content" and that raters form "a uniquely complex impression independently of the scale wordings." Part of this complex impression is related to raters' expectations of writers, often based on their own teaching and previous rating experience (see Weigle, 2002, for a discussion of this issue). Thus, raters may be influenced by their notions of the situations in which students would find themselves and may base their ratings in part on their intuitions about language issues that are problematic in content courses. This in turn may have aligned their scores more closely with instructor ratings.

Relationships between scores on TOEFL iBT essays and criterion variables. As for considerations of criterion-related validity, correlations between essay scores and other indicators of writing ability were generally moderate, whether they were scored by human raters or e-rater. These moderate correlations are not unlike those found in other criterion-related validity studies

(see, for example Kuncel, Hezlett, & Ones, 2001, for a meta-analysis of such studies of GRE Tests). They are also similar to or higher than those presented in Powers et al. (2000), which compared e-rater scores of GRE essays with a variety of other indicators. The correlations in that study ranged from .08 to .30 for a single human rater, .07 to .31 for two human raters, and .09 to .24 for e-rater. Possible explanations for the difference in magnitude of these correlations include improvements in e-rater since the Powers et al. study was written and difference in the writing constructs measured by the GRE and the TOEFL (Lee, 2006).

The highest correlations tended to be for global measures of global language proficiency rather than specific aspects of writing ability, suggesting that the TOEFL iBT independent task may be more useful as a measure of general language proficiency than of academic writing ability.

Research Question 2: Results and Discussion

Results

To answer Research Question 2 (regarding relationships among specific features analyzed by e-rater and indicators of writing ability), the eight e-rater feature scores were averaged across the two topics. Correlations were calculated between the averaged e-rater feature scores, the overall self-evaluation variables from the student survey, the overall evaluation of language proficiency from the instructor survey, and the writing sample scores. Results of these analyses are presented in Tables 17 to 19.

Table 17

Correlations of Averaged e-rater Feature Scores With Self-Evaluation of Language Skills

e-rater feature	Writing	Reading	Listening	Speaking
Vocabulary	.33**	.24**	.19**	.22**
Style	.27**	.29**	.19**	.22**
Usage	.27**	.25**	.25**	.19**
Grammar	.22**	.23**	.21**	.15**
Mechanics	.22**	.15**	.08	.08
Word Length	.19**	.15**	.08	.11*
Organization	.14**	.19**	.04	.07
Development	.13*	.14**	.14**	.14**

*p <. 05. **p < .01.

Table 18

Correlations of Averaged e-rater Feature Scores With Composite Instructor Evaluation of English Ability

e-rater feature	Subject instructor evaluation (n = 140)	English instructor evaluation (n = 144)
Usage	.37**	.25**
Grammar	.27**	.14
Style	.25**	.19*
Organization	.19*	–.02
Mechanics	.15	.17*
Development	.13	.20*
Vocabulary	.12	.19*
Word Length	.06	.21*

*p < .05. **p < .01.

Table 19

Correlations of Averaged e-rater Feature Scores With Averaged Content and Language Scores on Submitted Writing Samples

	Content average (two writing samples)	Language average (two writing samples)
Vocabulary	.37**	.44**
Grammar	.33**	.41**
Style	.31**	.35**
Mechanics	.26**	.31**
Word length	.25**	.29**
Usage	.24**	.29**
Development	.21**	.16**
Organization	.12*	.15**

*p < .05. **p < .01.

The variables that were most closely related to student self-perception of all four language skills were vocabulary, style, usage, and grammar. Other e-rater variables (mechanics, word length, and organization) were related to self-perceptions of writing and reading but not listening and speaking. All correlations were fairly low, however, with the exception of vocabulary, which had a correlation of .33 with self-perception of writing ability.

For the instructor ratings, the three ratings (writing, oral proficiency, and general English ability) were combined into a single scale, composite evaluation of English ability. Because of the differences between English and subject area instructors (discussed above), the correlations between this variable and the averaged e-rater feature scores were computed separately (see Table 18). These correlations show a different pattern from the self-evaluation variables above, although again the correlations are generally low. The usage and style variables were the only ones statistically significant for both groups of instructors. Grammar and organization were correlated with instructor evaluations for subject area instructors but not English instructors, and the other four variables were significantly, although only slightly, related to the English instructors' evaluations but not to the subject area instructors' evaluations.

As for correlations of e-rater feature scores and scores on course-related writing samples, the correlations in Table 19 are generally higher than those from the student and instructor surveys, particularly in the area of language. Vocabulary, grammar, and style have the highest correlations with these scores, with correlations above .30 for the language scores.

Discussion

In summary, the e-rater features that are consistently related to criterion variables tend to be those features related to linguistic accuracy (grammar, usage, style, and mechanics) rather than those related to discourse (organization and development). The highest correlations overall tended to be the correlations between these linguistic features as measured by e-rater and the ratings on the language scale of submitted writing samples. This result in particular lends support to the claim that the features evaluated by e-rater are similar to those used by human raters in judging the effective use of language.

One intriguing result is that vocabulary level was highly related to student self-evaluations and to ratings of writing samples, but not to instructor evaluations of language ability. This suggests that instructors may focus on language errors more than other factors when asked to evaluate their

students' language abilities in the abstract, whereas raters looking at specific writing samples take into account a wider variety of features of writing when giving scores.

Research Question 3: Results and Discussion

Results

To answer Research Question 3 (regarding consistency of scores across prompts), human and e-rater scores, as well as e-rater feature scores, were compared across the two writing prompts. Table 20 shows the descriptive statistics for the averaged human rater scores and e-rater scores for the two prompts. Recall that approximately half of the samples were rated by a group of four raters from ETS and the other half were rated by two raters at GSU; as noted previously, ratings were randomly assigned to either Rater 1 or Rater 2. Paired *t*-tests revealed that there were no significant differences across prompts for any rater.

To further investigate the effects of both prompt and rater on scores, a two-factor repeated-measures ANOVA was conducted with rater (Rater 1, Rater 2, e-rater) and topic (1 vs. 2) as independent variables. Results of this analysis are found in Table 21. As the table shows, the analysis revealed a main effect for Rater but no overall effect for Topic or Rater/Topic interaction. The rater effect can be seen in Figure 1: E-rater scores are lower than either of the two human ratings. However, the effect is small (partial $eta^2 = .03$).

Table 20

Descriptive Statistics and Paired t-Tests for Human and e-rater Raters Across Prompts

	N	Mean	*SD*	*t*	*p*
T1 average human score	376	3.36	0.81	-.75	.45
T2 average human score	376	3.38	0.89		
T1 e-rater score	377	3.24	0.91	-.24	.79
T2 e-rater score	377	3.24	0.94		

Note. *t*-test results refer to comparisons of ratings by human raters and e-rater, respectively, across the two prompts

Table 21

Repeated-Measures ANOVA for Topic and Rater on Scores Using Individual Raters

Source	Type III *SS*	*Df*	*MS*	*F*	Sig.	Partial eta^2
Topic	0.33	1	0.33	1.21	.27	.003
Error (Topic)	102.91	375	0.27			
Rater	8.51	2	4.25	11.53	.00	.030
Error (Rater)	276.51	750	0.37			
Topic * Rater	0.19	2	0.09	0.36	.70	.001
Error (Topic * Rater)	196.16	750	0.26			

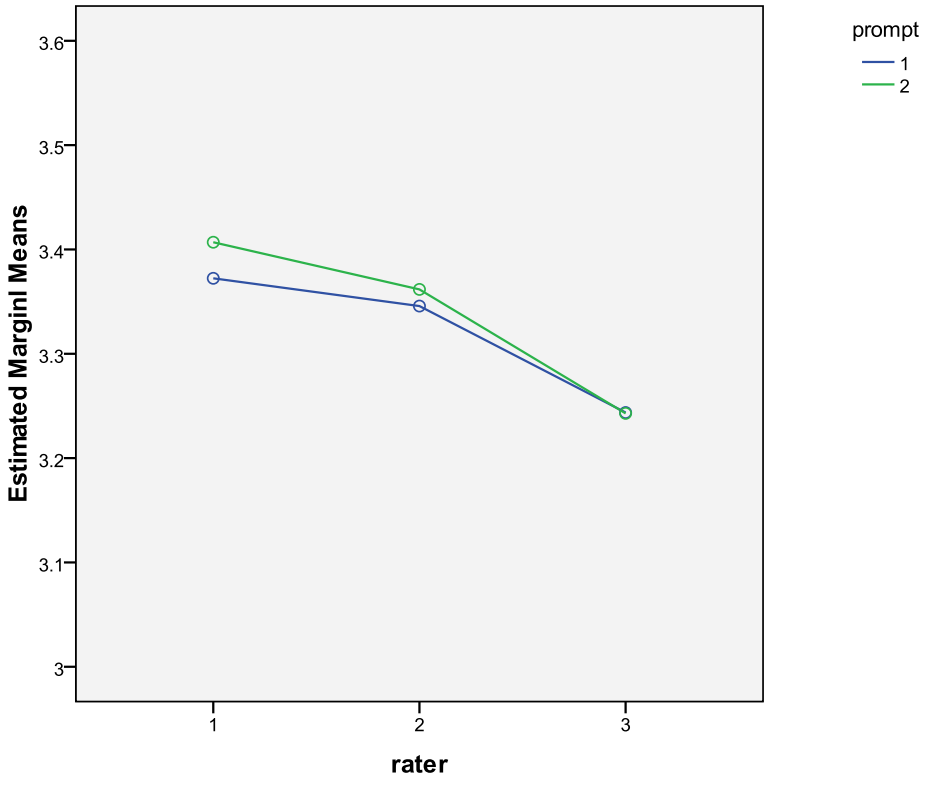


Figure 1. **Plot of rater-by-topic interaction (single human raters vs. e-rater).**

Table 22 shows the descriptive statistics, correlations, and results of paired *t*-tests of the e-rater features between the two topics. The features are presented in descending order of Pearson correlations between topics. As the table shows, the feature scores differed across topics in several ways. The mean scores were significantly different across topics for all features except grammar and organization, with three features (word length, vocabulary, and style) having higher means on Topic 1 and others (development, mechanics, and usage) having higher means on Topic 2 (significantly higher mean scores are in boldface in the table). The features that were most strongly correlated across topics were mechanics (.71), vocabulary (.63), and organization (.60), with the lowest correlations for usage (.33) and grammar (.34).

Table 22

Descriptive Statistics, Correlations, and Results of Paired t-Tests Across Prompts of e-rater Features (N = 377)

		Topic 1		Topic 2				
	r	Mean	*SD*	Mean	*SD*	*t*	*p*	Cohen's *d*
Mechanics	.71	-0.16	0.07	**-0.13**	0.07	-9.01	.00	-0.56
Vocabulary	.63	**-62.25**	2.92	-64.41	2.55	17.66	.00	0.92
Organization	.60	1.92	0.37	1.88	0.37	1.75	.08	0.12
Development	.56	3.71	0.36	**3.84**	0.37	-7.45	.00	-0.38
Word Length	.49	**4.59**	0.24	4.23	0.24	28.31	.00	1.49
Style	.44	**-0.21**	0.13	-0.25	0.13	5.16	.00	0.29
Grammar	.34	-0.10	0.03	-0.09	0.03	-1.19	.24	-0.06
Usage	.33	-0.10	0.04	**-0.09**	0.04	-2.49	.01	- 0.21

Note. Significantly higher mean scores are in boldface.

Discussion

One criticism frequently leveled at large-scale writing testing programs is that decisions are often made on the basis of a single essay written on a single topic. ETS has responded to such criticisms in TOEFL iBT by including two different writing tasks (one independent and one integrated); however, test takers still only respond to a single independent writing task, and thus it is important to demonstrate that the specific prompt given to the test taker does not affect the outcome of the test. The last research question addresses this issue.

The results of this analysis show that e-rater is generally more consistent across the two prompts than individual human raters. This consistency is a potentially powerful argument in favor of using e-rater operationally to eliminate or reduce human bias related to different writing prompts. However, in practice differences in individual raters are mitigated by double-scoring and using a large number of different raters, so that no individual rater bias has a serious detrimental effect on test results. In this study rater bias may have been magnified by the use of a small number of human raters who were trained but not highly experienced in rating TOEFL essays. Human raters tended to rate essays slightly higher than e-rater did, but the effect size was small and of little practical significance; furthermore, this result may have been due to the specific raters used for the study.

However, even though the overall e-rater scores were consistent across the two topics, the specific e-rater feature scores did not show the same consistency. There were significant differences in six of the eight feature scores across the two topics; the largest differences were found in the vocabulary and word length scores, with higher scores on Topic 1 (importance of appearance) than Topic 2 (planning for the future). At the same time the vocabulary score was highly correlated across the two topics, which indicates that the prompts rank-ordered students in terms of their vocabulary usage in similar ways. An analysis of the 50 most frequently used words among the student responses reveals that more words of over six letters were used frequently for Topic 1 (appearance, fashion, personal, emphasis, and important) than for Topic 2 (planning and carefully); since several of these words appear in the respective prompts, the differences in vocabulary and word length scores may simply be an artifact of the prompt wording.

The differences in e-rater feature scores could be explored more thoroughly through a textual analysis of TOEFL essays. For example, it may be the case that Topic 2 elicited more personal narratives (and thus more instances of past tense) in support of arguments, while Topic 1 elicited supporting arguments in the form of general truth statements in the present tense, resulting in a lower correlation between the grammar scores on the two topics (see Biber, 1988, for a more complete discussion of the relationship between discourse genres and specific linguistic structures).

The data presented here are also similar to data presented in Attali and Burstein (2006), who found the true-score correlation between human and machine scores to be .97, a result replicated in this study. On the other hand the alternate-forms reliability of individual e-rater features presented by Attali and Burstein from essays written by 6th through 12th graders differs from the reliabilities found in this study (see Table 23). In particular the mechanics reliability coefficient in the current study is quite a bit higher than the same coefficient in the Attali and Burstein study, while the grammar and usage reliabilities are lower. The range of reliabilities is also greater. These differences may be due to the fact that different populations were used in the two studies—Attali and Burstein's data came from middle and secondary students, including both native and nonnative speakers, while this study's data came from university students, all of whom are nonnative speakers of English.

Table 23

Comparison of e-rater Feature Correlations Across Alternate Forms of Writing Tests

Feature	Present study	Attali & Burstein (2006)
Mechanics	.71	.46
Vocabulary	.63	.44
Organization	.60	.48
Development	.56	.36
Word Length	.49	.47
Style	.44	.43
Grammar	.34	.45
Usage	.33	.45

Implications and Future Directions

This study adds to the growing literature related to the validation and use of e-rater for TOEFL essays (e.g., Attali, 2007, 2008; Attali & Burstein, 2006; Chodorow & Burstein, 2004; Enright & Quinlan, 2008; Lee et al., 2008). In terms of the validity argument for the TOEFL outlined by Chapelle et al. (2008), the study provides evidence that support the inferences of generalization (across tasks and raters) and extrapolation to other criteria of writing ability in academic contexts. As for the inference of utilization--that is, the inference that test scores obtained in part through the use of e-rater are clearly interpretable and that the test will have a positive influence on English teaching—perhaps one of the most serious deterrents to the operational use of automated scores in high-stakes assessment is the opposition to machine scoring by groups of stakeholders, especially teachers of writing and perhaps students themselves. Evidence of the comparability of e-rater and human scores in terms of how they relate to nontest indicators of writing ability may promote acceptance of automated scoring, especially if it can be argued that automated scoring would help to contain the costs and resources needed to score the test and mitigate the need to pass additional costs on to test takers, thus potentially increasing access to the test. It may be that stakeholders are more willing to accept automated scoring of second-language proficiency tests than of tests intended primarily for first-language writers, since quantifiable sentence-level aspects of texts are more intuitively related to language proficiency than are concerns of voice, audience awareness, and the ability to make a persuasive argument, notions that are of paramount importance to composition teachers.

At the same time the study results suggest that e-rater cannot duplicate human ratings, and there are still some differences between e-rater and human scores. This may even be a comforting observation to writing teachers, since it emphasizes the fact that writing is primarily a means of communicating between people, not a collection of measurable features of text. Presenting e-rater as a tool to help streamline the process of making decisions on the basis of test scores rather than as a substitute for human judgment may help allay the fears of those who object to machine scoring of writing.

There are a number of limitations to this study, and a few of them will be mentioned here. A number of sources of unreliability may have affected the correlations presented. The human ratings on TOEFL essays may have been affected by the fact that the raters hired by the author were not as extensively trained or experienced as the ETS raters. The writing samples submitted

by students were also of such varied nature that consistent scoring may have been a problem. In addition, those writing samples were written in very different contexts, and participants had access to varying levels of assistance and feedback, so they are not all strictly products of the individuals who turned them in. The instructors who agreed to participate in the study also were not a random sample; students may have chosen those teachers they thought would give positive reports about their writing ability, or instructors who had particularly strong feelings about their students' language proficiency may have been more willing to take the survey. All of these variables may have affected the study results.

Nevertheless, the data collected for this study represent a wealth of information about the relationship between writing test scores and the role of writing in student success, and additional analyses tangential to the focus of this report may be carried out. One analysis that might shed light on some of the differences across topics would be a many-faceted Rasch analysis using the FACETS software (Linacre, 2010; see also Myford & Wolfe, 2003, 2004, for details of this method of analysis), which can be used to estimate rater severity and task difficulty on the same linear scale, allowing investigation of questions such as whether specific raters judge essays on certain topics more severely than others. This analysis could provide more detailed information about rater bias, and along with e-rater feature scores could complement recent research on the factors that influence rater behavior (e.g., Eckes, 2008).

Other questions less directly related to the focus of this report concern the relationship between students' perceptions of their abilities and their chosen field (do students at the same level of proficiency but in different areas of study, such as science vs. business, perceive their language needs differently?) or between a student's self-assessment of his or her own ability and the perceptions of that student's instructor. The corpus of writing samples collected for this study could also be useful in exploring questions about the role of writing in academic life. For example, information was gathered on the types of help and feedback students received on the writing samples they submitted; these data have not yet been analyzed and might shed light on the academic lives of nonnative speakers in terms of what resources and support they find useful in their writing. The analysis of writing samples in terms of register, cognitive demands, and use of sources, while ultimately not proving overly informative for the present study, may have implications for the study of the types of writing that students need to do for their coursework and may in fact be more useful for subsequent analyses not related specifically to the TOEFL.

In conclusion, the findings presented in this report highlight the complex nature of writing, the relationships between writing on a test and writing in academia, and the use of different testing and scoring methods to make judgments about students' readiness to participate fully in academic life. The use of automated scoring in conjunction with human scoring to make the process of assessing writing more efficient and potentially more reliable is just one of many factors that may affect the outcome of assessments. There is no single ideal testing format or scoring procedure, but e-rater certainly holds promise as an additional tool in the language tester's toolkit.

References

Allwright, J., & Banerjee, J. (1997). *Investigating the accuracy of admissions criteria: A case study in a British university* (CRILE Occasional Report No.7). United Kingdom: Lancaster University, Centre for Research in Language Education.

Anson, C. (2006). Can't touch this: Reflections on the servitude of computers as readers. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of human essays* (pp. 38–56). Logan: Utah State University Press.

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS Research Report No. RR-07-21). Princeton, NJ: ETS.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Ayres, J. B., & Peters, R. M. (1977). Predictive validity of the test of English as a foreign language for Asian graduate students in engineering, chemistry or mathematics. *Educational and Psychological Measurement, 37*(2), 461–463.

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

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment*, *6*(1). Available from http://www.jtla.org

Biber, D. (1988). *Variation across speech and writing.* United Kingdom: Cambridge University Press.

Breland, H. M., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Report No. 99-3; GRE Board Research Report No. 96-12R; ETS RR-99-03). New York, NY: The College Board.

Burstein, J. (2002). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In M. B. Olson (Ed.), Proceedings *of the ACL99 Workshop on computer-mediated language assessment and evaluation of natural language processing* (pp. 68-75). College Park, MD: Association of Computational Linguistics.

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). Criterion^{TM} online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence,* Acapulco, Mexico. Retrieved from http://www.ets.org/Media/Research/pdf/erater_iaai03_burstein.pdf

Chapelle, C., Enright, M., & Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright, & J. Jameison (Eds.), *Building a validity argument for the Test of English as a Foreign Language^{TM}*. New York, NY: Routledge

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

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL Research Report No. 73; ETS RR-04-04). Princeton, NJ: ETS.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of grammatical error detection systems for English language learners: Feedback and assessment. *Language Testing, 27*(3), 419–436.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Condon, W. (2006). Why less is not more: What we lose by letting a computer score writing samples. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of human essays* (pp. 211–220). Logan: Utah State University Press.

Conference on College Composition and Communication. (2004, February 25). *CCCC position statement on teaching, learning, and assessing writing in digital environments*. Retrieved from http://www.ncte.org/cccc/resources/positions/digitalenvironments

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39*(1), 81–141.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155–185.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Lawrence Erlbaum Associates.

Enright, M., & Quinlan, T. (2010). Using e-rater® to score essays written by English language learners: A complement to human judgment. *Language Testing 27*(3), 317–334.

Golub-Smith, M. L., Reese, C. M., & Steinhaus, K. (1993). *Topic and topic type comparability on the Test of Written English*™ (TOEFL Research Report No. 42; ETS RR-93-10). Princeton, NJ: ETS.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

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

Kane, M. (2001). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–35.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (TOEFL Research Report No. 59; ETS RR-98-06). Princeton, NJ: ETS.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations®: Implications for graduate student selection and performance. *Psychological Bulletin, 127*(1), 162–181.

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

Lee, Y.-W. (2006). *Variability and validity of automated essay scores for TOEFL iBT: Generic, hybrid, and prompt-specific models.* Unpublished manuscript. Princeton, NJ: ETS.

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL® CBT essays: Scores from humans and e-rater®* (TOEFL Research Report No. RR-81; ETS RR-08-01). Princeton, NJ: ETS.

Light, R.L., Xu, M., & Mossop, J. (1987*).* English proficiency and academic performance of international students. *TESOL Quarterly, 21*(2), 251–261.

Linacre, J. M. (2010). Facets Rasch measurement computer program, version 3.67.1 [Computer software]. Chicago, IL: Winsteps.com.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246–276.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19,* 477-496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-62.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4,* 386-422.

Myford, C. M., & Wolfe, E. W . (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5,* 189-227.

Neal, M. E. (1998). *The predictive validity of the GRE and TOEFL exams with GGPA as the* Criterion *for international graduate students in science and engineering*. (ERIC Document Reproduction Service No. ED 424294)

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*(1), 238–243.

Page, E. B. (2003). Project Essay Grade: PEG. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–56). Mahwah, NJ: Lawrence Erlbaum Associates.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K**.** (2000). *Comparing the validity of automated and human essay scores* (GRE Board Research Report No. 98-08a; ETS RR-00-10). Princeton, NJ: ETS.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the teaching of English, 26,* 108–122.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46,* 137–174.

Simner, L. (1999). Reply to the universities' reaction to the Canadian Psychological Association's position statement on the TOEFL. *European Journal of Psychological Assessment, 15*(3), 284–294.

Urry, H. L. (2003, March 7). Re: test to compare correlations from the SAME sample? Message posted to http://www.listserv.uga.edu/cgi-bin/wa?A2=ind0303&L=spssx-l&P=9391.

Weigle, S. (2002). *Assessing writing.* Cambridge: United Kingdom: Cambridge University Press.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning and Technology, 8*(1), 53–65.

Xu, M. (1991). The impact of English-language proficiency on international graduate students' perceived academic difficulty. *Research in Higher Education, 32*, 557–570.

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

Yule, G., & Hoffman, P. (1990). Predicting success for International Teaching Assistants in a U.S. university. *TESOL Quarterly, 24*(2), 227–243.

Notes

[1] The SPSS syntax used for these calculations was retrieved on September 1, 2009, from http://ssc.utexas.edu/consulting/answers/general/gen28.html

[2] *t*-test results are shown for Sample 1 only; similar results were found for Sample 2.

List of Appendices

Appendix A

Student Survey

1. Introduction

We would like to ask you a few questions about your English ability. This survey will only take five to ten minutes of your time.

*** 1. Please enter the four-digit code you were given today.**

2. English language ability

How do you assess your ability to use English in your major subject? If you do not yet have a major subject, please answer each question with reference to the courses in the subject area you know most about.

For each question below, check the phrase which describes your ability best.

2. How well do you write about your subject area in English?

◯ Not well enough to survive ◯ Just well enough to survive ◯ Reasonably well ◯ Fully able to meet all my study needs

3. How well do you read about your subject area in English?

◯ Not well enough to survive ◯ Just well enough to survive ◯ Reasonably well ◯ Fully able to meet all my study needs

4. How well do you understand what your instructors say about your subject area in English?

◯ Not well enough to survive ◯ Just well enough to survive ◯ Reasonably well ◯ Fully able to meet all my study needs

5. How well do you speak about your subject area in English?

◯ Not well enough to survive ◯ Just well enough to survive ◯ Reasonably well ◯ Fully able to meet all my study needs

6. How would you compare your ability to use English for your coursework with your ability to use English outside of school (e.g. with friends, to take care of errands, etc.)

◯ School-related English is much better ◯ School-related English is somewhat better ◯ about the same ◯ Everyday English is somewhat better ◯ Everday English is much better

7. Any other comments about your English language ability:

3. Writing in English

8. For this question, please think about the WRITING you need to do for your courses. Please respond with reference to the course(s) in which you have the MOST difficulties.

The following are areas where some students experience problems with writing in their course work, even if they are getting good grades. Please indicate how often you have problems with each area below.

How much of a problem do you find each of the following?

	A very serious problem	A somewhat serious problem	Occasionally a problem	Not a problem at all	Not Applicable
Understanding what the writing assignment requires me to do	◯	◯	◯	◯	◯
Organizing my ideas	◯	◯	◯	◯	◯
Expressing my ideas clearly	◯	◯	◯	◯	◯
Using appropriate vocabulary related to the subject	◯	◯	◯	◯	◯
Using appropriate general vocabulary	◯	◯	◯	◯	◯
Using English grammar correctly	◯	◯	◯	◯	◯
Knowing how to format my papers correctly	◯	◯	◯	◯	◯
Doing my best writing within the amount of time given	◯	◯	◯	◯	◯
Writing on a computer	◯	◯	◯	◯	◯

9. Please comment on any other problems that you might have in writing:

4. Other language-related difficulties

10. The following are areas where some students experience other language-related problems in their course work, even if they are getting good grades.

Please answer these questions with reference to the course(s) in which you have the MOST difficulties.

How much of a problem do you find each of the following?

	A very serious problem	A somewhat serious problem	Occasionally a problem	No problem at all	Not Applicable
Understanding my instructor's lectures in class	◯	◯	◯	◯	◯
Understanding my classmates	◯	◯	◯	◯	◯
Participating in class discussions	◯	◯	◯	◯	◯
Understanding the course readings	◯	◯	◯	◯	◯
Reading fast enough to keep up with the amount of reading required	◯	◯	◯	◯	◯

11. Please comment on any other language-related problems that you are experiencing:

5. Difficulties due to non-language-related factors

12. Many students have problems that are not related to language, even if they are getting good grades. Here is a list of problems that students like yourself may encounter. Please answer this question with reference to the course(s) in which you have the MOST difficulties.

How much of a problem is each one for you?

	A very serious problem	A somewhat serious problem	Occasionally a problem	No problem at all	Not applicable
Understanding the subject matter and concepts related to the subject matter	◯	◯	◯	◯	◯
Getting used to the way classes are taught in the U.S.	◯	◯	◯	◯	◯
Knowing how to get help with coursework when I need it	◯	◯	◯	◯	◯
Organizing my time to get all my school work done	◯	◯	◯	◯	◯

13. Please comment on any other difficulties you may be having:

Appendix B

Factor Analysis of Student Survey Variables

Table B1

Unrotated Factor Matrix

	Factor 1	*Factor 2*	*Factor 3*
Understanding what the writing assignment requires me to do	.685	.037	.162
Organizing my ideas	.647	.261	.176
Expressing my ideas clearly	.712	.200	-.196
Using appropriate vocabulary related to the subject	.678	.144	-.456
Using appropriate general vocabulary	.700	.105	-.401
Using English grammar correctly	.587	.387	-.282
Knowing how to format my papers correctly	.621	.342	.041
Doing my best writing within the amount of time given	.548	.271	-.145
Writing on a computer	.371	.365	.151
Understanding my instructor's lectures in class	.654	-.499	.021
Understanding my classmates	.568	-.522	-.179
Participating in class discussions	.566	-.474	-.191
Understanding the course readings	.664	-.257	.073
Reading fast enough to keep up with the amount of reading required	.651	-.215	.026
Understanding the subject matter and concepts related to the subject matter	.711	-.130	.191
Getting used to the way classes are taught in the U.S.	.572	-.038	.331
Knowing how to get help with coursework when I need it	.672	.052	.298
Organizing my time to get all my school work done	.530	.092	.514

Table B2

Rotated Factor Matrix

	Factor 1	*Factor 2*	*Factor 3*
Using appropriate vocabulary related to the subject (WP4)	0.75		
Using English grammar correctly (WP6)	0.73		
Using appropriate general vocabulary (WP5)	0.71		
Expressing my ideas clearly (WP3)	0.65		
Doing my best writing within the amount of time given (WP8)	0.56		
Knowing how to format my papers correctly (WP7)	0.53		
Understanding my classmates (LP2)		0.77	
Understanding my instructor's lectures in class (LP1)		0.77	
Participating in class discussions (LP3)		0.74	
Understanding the course readings (LP4)		0.57	
Reading fast enough to keep up with the amount of reading required (LP5)		0.54	
Organizing my time to get all my school work done (OP4)			0.73
Knowing how to get help with coursework when I need it (OP3)			0.63
Getting used to the way classes are taught in the U.S. (OP2)			0.58
Organizing my ideas (WP2)			0.57
Understanding what the writing assignment requires me to do (WP1)			0.53
Understanding the subject matter and concepts related to the subject matter (OP1)			0.52
Writing on a computer (WP9)			0.42

Appendix C

Instructor Survey

1. English ability for academic performance

Thank you for agreeing to participate in our study. We would like to ask you a few questions about your perceptions of this student's English ability and his/her performance in your class. **The survey should take five to ten minutes and your responses will be kept completely confidential.**

Information from the study will be used to improve the scoring of writing tests and to help make informed decisions about whether or not it is appropriate to use computers to score essays written by non-native speakers of English.

2. Untitled Page

*** 1. Please enter the four-digit Instructor Code from the email that was sent to you.**

*** 2. Please enter the four-digit Student Code from the email that was sent to you.**

3. Please enter the course number(s) and title(s) in which the student is/was enrolled, e.g., "English 1101: Introduction to Composition"

4. How would you rate this student's overall academic performance in the course?

- ◯ The student is not performing/performed well enough to pass
- ◯ The student is/was on the borderline.
- ◯ The student is performing/performed well.
- ◯ The student is performing/performed at an exceptionally high level.
- ◯ No opportunity to judge

3. Ability in English

5. How would you rate this student's English proficiency in the following areas?

	Very limited user	Modest user	Good user	Excellent user	no opportunity to judge
writing ability for academic purposes	◯	◯	◯	◯	◯
oral proficiency (speaking and listening) for academic purposes	◯	◯	◯	◯	◯
overall English ability for academic purposes	◯	◯	◯	◯	◯

6. Comments on student's overall English language ability

4. Impact of student's English on performance

7. Some students have difficulties in courses that can be attributed to inadequate English, even if they are performing satisfactorily in class. Please respond to the items below even if the student is performing well in your class.

In my view, the student is having/has had serious problems...

	Very often	Often	Sometimes	Rarely	Never	I can't judge
Understanding the course readings	◯	◯	◯	◯	◯	◯
Understanding the requirements of writing assignments	◯	◯	◯	◯	◯	◯
Organizing ideas in writing	◯	◯	◯	◯	◯	◯
Expressing ideas clearly in writing	◯	◯	◯	◯	◯	◯
Using appropriate vocabulary in writing	◯	◯	◯	◯	◯	◯
Using English grammar correctly in writing	◯	◯	◯	◯	◯	◯
Understanding lectures and directions in class	◯	◯	◯	◯	◯	◯
Participating in class discussions	◯	◯	◯	◯	◯	◯
Asking for help when necessary	◯	◯	◯	◯	◯	◯

5. Impact of nonlinguistic factors on performance

8. Do you feel that this student has difficulties in your class are not related to English language ability? Please answer each item even if the student is performing satisfactorily in your class.

In my view, the student is having/has had serious problems because of...

	Very often	Often	Sometimes	Rarely	Never	I can't judge
Lack of familiarity with the university culture	◯	◯	◯	◯	◯	◯
Lack of appropriate background for the course	◯	◯	◯	◯	◯	◯
Difficulties understanding concepts related to the subject matter	◯	◯	◯	◯	◯	◯
Inadequate study skills	◯	◯	◯	◯	◯	◯
Other nonlinguistic factors	◯	◯	◯	◯	◯	◯

9. Please list other nonlinguistic factors and any comments.

10. May we contact you for additional information about your responses if necessary?

◯ Yes ◯ No

11. If yes, please provide a phone number and/or email address.

56

12. Would you like to receive information about the study when it is finished?

◯ Yes ◯ No

13. If yes, please provide a mailing address and/or email address.

6. Untitled Page

Thank you for your assistance!

Appendix D

Factor Analysis of Instructor Survey Variables

Table D1

Unrotated Factor Matrix

	Component	
	1	2
Understanding the course readings	.844	.309
Understanding the requirements of writing assignments	.878	.289
Organizing ideas in writing	.875	.255
Expressing ideas clearly in writing	.860	.267
Using appropriate vocabulary in writing	.880	.207
Using English grammar correctly in writing	.846	.224
Understanding lectures and directions in class	.864	.240
Participating in class discussions	.767	.165
Asking for help when necessary	.775	.317
Lack of familiarity with the university culture	.226	.765
Lack of appropriate background for the course	.257	.866
Difficulties understanding concepts related to the subject matter	.249	.827
Inadequate study skills	.237	.831
Other nonlinguistic factors	.231	.826

Table D2

Rotated Factor Matrix

	Component	
	1	2
Understanding the course readings	.86	
Understanding the requirements of writing assignments	.92	
Organizing ideas in writing	.89	
Expressing ideas clearly in writing	.87	
Using appropriate vocabulary in writing	.90	
Using English grammar correctly in writing	.86	
Understanding lectures and directions in class	.83	
Participating in class discussions	.77	
Asking for help when necessary	.79	
Lack of familiarity with the university culture		.81
Lack of appropriate background for the course		.91
Difficulties understanding concepts related to the subject matter		.87
Inadequate study skills		.85
Other nonlinguistic factors		.86

Appendix E

Participant Information Form

Research Code: _____

Participant Information Form

Background Information

1. Family Name: _____
2. First Name: _____
3. Native Language: _____
4. Year of Birth: _____
5. Gender: Female : □ Male: □
6. Phone number: _____
7. Email: _____
8. Major: _____
9. Status: Graduate: □ Undergraduate: □

English Usage

10. For how many years have you studied English? _____
11. How old were you when you started to study English? _____
12. How long have you lived in an English-speaking country? _____
13. What English classes are you studying now? (Provide class numbers and names.)

_____

14. Have you ever taken the TOEFL test? Yes □ No □
15. If yes:

    What was your score? _____

    When did you take the test? _____
16. How many hours per week do you spend writing in English outside class? _____

Writing Sample #1

17. What class was this paper written for? _____
18. When was it written? _____
19. When you wrote this paper, did you get any help with the English? Check all that apply.

    Dictionary: □

    Grammar book: □

    Help from the teacher: □

    Help from a classmate: □

Help from a tutor: □

Other: □ Describe: _____

20. If "3" represents your typical writing, how would you rate this paper? 4 = Best, 1 = Worst

4 □ 3 □ 2 □ 1 □

Writing Sample #2

21. What class was this paper written for? _____
22. When was it written? _____
23. When you wrote this paper, did you get any help with the English? Check all that apply.

Dictionary: □

Grammar book: □

Help from the teacher: □

Help from a classmate: □

Help from a tutor: □

Other: □ Describe: _____

24. If "3" represents your typical writing, how would you rate this paper? 4 = Best, 1 = Worst

4 □ 3 □ 2 □ 1 □

Reference #1: Contact Information

25. What is name of your reference?_____
26. How do you know this person? (For example, which course did you take with him/her?)

_____

27. Contact's telephone number: _____
28. Contact's email address: _____

Reference #2: Contact Information

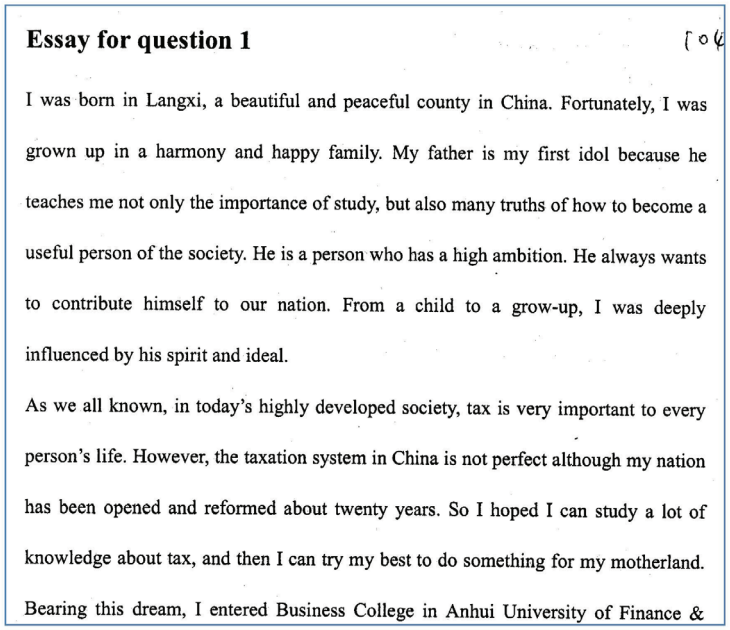
29. What is name of your reference?_____
30. How do you know this person? (For example, which course did you take with him/her?)

_____

31. Contact's telephone number: _____
32. Contact's email address: _____

Appendix F

Samples of Submitted Student Writing

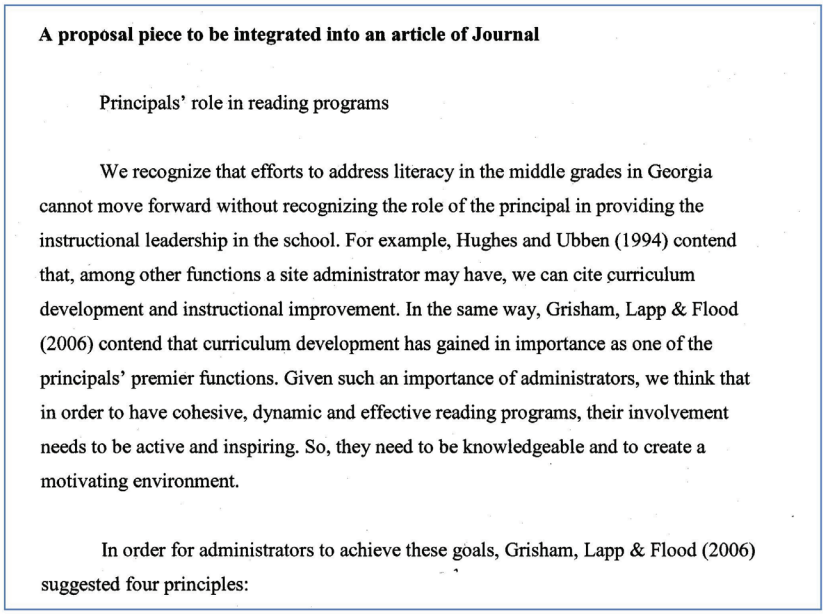


Essay for question 1

I was born in Langxi, a beautiful and peaceful county in China. Fortunately, I was grown up in a harmony and happy family. My father is my first idol because he teaches me not only the importance of study, but also many truths of how to become a useful person of the society. He is a person who has a high ambition. He always wants to contribute himself to our nation. From a child to a grow-up, I was deeply influenced by his spirit and ideal.

As we all known, in today's highly developed society, tax is very important to every person's life. However, the taxation system in China is not perfect although my nation has been opened and reformed about twenty years. So I hoped I can study a lot of knowledge about tax, and then I can try my best to do something for my motherland. Bearing this dream, I entered Business College in Anhui University of Finance &

Figure F1. **Sample 1.**



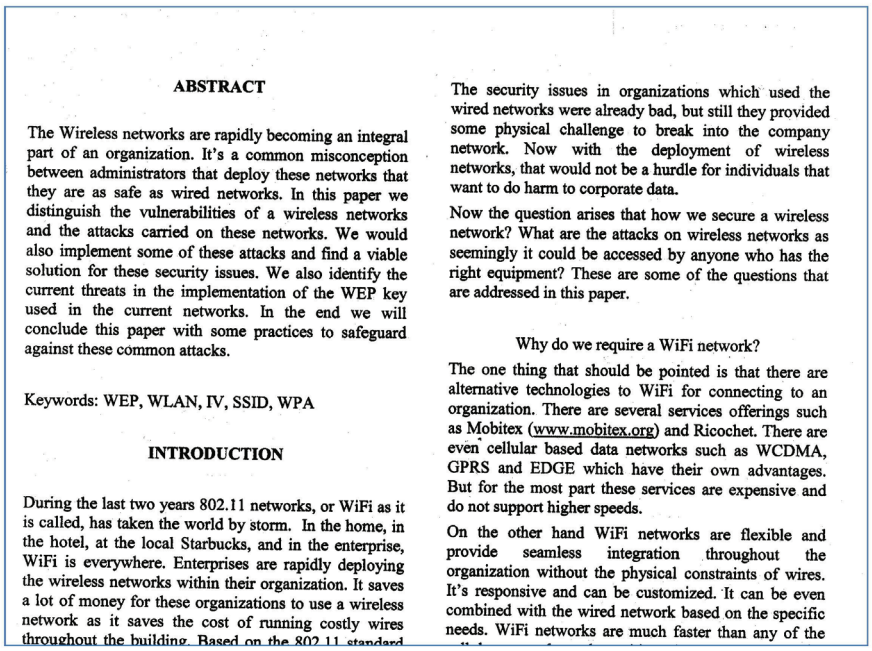
A proposal piece to be integrated into an article of Journal

Principals' role in reading programs

We recognize that efforts to address literacy in the middle grades in Georgia cannot move forward without recognizing the role of the principal in providing the instructional leadership in the school. For example, Hughes and Ubben (1994) contend that, among other functions a site administrator may have, we can cite curriculum development and instructional improvement. In the same way, Grisham, Lapp & Flood (2006) contend that curriculum development has gained in importance as one of the principals' premier functions. Given such an importance of administrators, we think that in order to have cohesive, dynamic and effective reading programs, their involvement needs to be active and inspiring. So, they need to be knowledgeable and to create a motivating environment.

In order for administrators to achieve these goals, Grisham, Lapp & Flood (2006) suggested four principles:

Figure F2. **Sample 2.**



ABSTRACT

The Wireless networks are rapidly becoming an integral part of an organization. It's a common misconception between administrators that deploy these networks that they are as safe as wired networks. In this paper we distinguish the vulnerabilities of a wireless networks and the attacks carried on these networks. We would also implement some of these attacks and find a viable solution for these security issues. We also identify the current threats in the implementation of the WEP key used in the current networks. In the end we will conclude this paper with some practices to safeguard against these common attacks.

Keywords: WEP, WLAN, IV, SSID, WPA

INTRODUCTION

During the last two years 802.11 networks, or WiFi as it is called, has taken the world by storm. In the home, in the hotel, at the local Starbucks, and in the enterprise, WiFi is everywhere. Enterprises are rapidly deploying the wireless networks within their organization. It saves a lot of money for these organizations to use a wireless network as it saves the cost of running costly wires throughout the building. Based on the 802.11 standard

The security issues in organizations which used the wired networks were already bad, but still they provided some physical challenge to break into the company network. Now with the deployment of wireless networks, that would not be a hurdle for individuals that want to do harm to corporate data.

Now the question arises that how we secure a wireless network? What are the attacks on wireless networks as seemingly it could be accessed by anyone who has the right equipment? These are some of the questions that are addressed in this paper.

Why do we require a WiFi network?

The one thing that should be pointed is that there are alternative technologies to WiFi for connecting to an organization. There are several services offerings such as Mobitex (www.mobitex.org) and Ricochet. There are even cellular based data networks such as WCDMA, GPRS and EDGE which have their own advantages. But for the most part these services are expensive and do not support higher speeds.

On the other hand WiFi networks are flexible and provide seamless integration throughout the organization without the physical constraints of wires. It's responsive and can be customized. It can be even combined with the wired network based on the specific needs. WiFi networks are much faster than any of the

Figure F3. **Sample 3.**

104

Modeling Fluids by SAFT-VR and SAFT-1

I. Introduction

To modeling the thermodynamic properties of fluids, there are many approaches. A classic one is the statistical associating fluid theory (SAFT)[1] model, which is based on the thermodynamic perturbation theory. After that, a lot of other models related to SAFT were created, such as SAFT-HS, SAFT-LJ, SAFT-VR[2], and SAFT-1[3]. In these models, SAFT-VR and SAFT-1 are two good ones. By SAFT-VR, we treat molecules as chains of segments with an attractive potential of variable range. And by SAFT-1, we treat molecules as chains of square-well segments.

When we modeling fluids by SAFT based approaches, we consider the fluids by several parameters: σ, the hard-sphere diameter; ε, the depth of square well potential; λ, the cutoff radius for the association interactions; m, the united atom, $m=(C-1)/3+1$, where C is the numbers of carbon; and ε_{ab} the association energy.

In this project, I modeled methane, propane, and water by SAFT-VR and SAFT-1. In these three fluids, methane is a simplest one: m=1, ε_{ab}=0, no associating attraction. For propane, m=1.667, ε_{ab}=0. And for water, m=1, $\varepsilon_{ab}\neq 0$. Also, the results got by SAFT-VR and SAFT-1 are compared.

Figure F4. **Sample 4.**

Appendix G

TOEFL Independent Writing Task Scoring Guide

5 — An essay at this level largely accomplishes all of the following:

- effectively addresses the topic and task
- is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details
- displays unity, progression, and coherence
- displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors

4 — An essay at this level largely accomplishes all of the following:

- addresses the topic and task well, though some points may not be fully elaborated
- is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details
- displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections
- displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning

3 — An **essay** at this level is marked by one or more of the following:

- addresses the topic and task using somewhat developed explanations, exemplifications, and/or details
- displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
- may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning
- may display accurate but limited range of syntactic structures and vocabulary

2 — An essay at this level may reveal one or more of the following weaknesses:

- limited development in response to the topic and task
- inadequate organization or connection of ideas
- inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

1 — 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, or questionable responsiveness to the task
- serious and frequent errors in sentence structure or usage

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)

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

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