ETS TOEFL

Quality Beyond Measure.

TOEFL iBT™ Research

Insight

Series I, Volume 2

TOEFL Research

ETS — *Listening. Learning. Leading.®*

Foreword

We are very excited to announce the TOEFL iBT™ Research Insight Series, a bimonthly publication to make important research on the TOEFL iBT available to all test score users in a user-friendly format.

The TOEFL iBT test is the most widely accepted English language assessment, used for admissions purposes in more than 130 countries including the United Kingdom, Canada, Australia, New Zealand and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the TOEFL iBT test , was launched in 2005. It contains a number of innovative design features, including the use of integrated tasks that engage multiple language skills to simulate language use in academic settings, and the use of test materials that reflect the reading and listening demands of real-world academic environments.

At ETS we understand that you use TOEFL iBT test scores to help make important decisions about your students, and we would like to keep you up-to-date about the research results that assure the quality of these scores. Through the TOEFL iBT Research Insight Series we wish to both communicate to the institutions and English teachers who use the TOEFL iBT test scores the strong research and development base that underlies the TOEFL iBT test, and demonstrate our strong, continued commitment to research.

We hope you will find this series relevant, informative and useful. We welcome your comments and suggestions about how to make it a better resource for you.

Ida Lawrence
Senior Vice President
Research & Development Division
Educational Testing Service

Preface

Since the 1970's, the TOEFL test has had a rigorous, productive and far-ranging research program. But why should test score users care about the research base for a test? In short, because it is only through a rigorous program of research that a testing company can demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. This is why ETS has made the establishment of a strong research base a consistent feature of the evolution of the TOEFL test.

The TOEFL test is developed and supported by a world-class team of test developers, educational measurement specialists, statisticians and researchers. Our test developers have advanced degrees in such fields as English, language education and linguistics. They also possess extensive international experience, having taught English in Africa, Asia, Europe, North America and South America. Our research, measurement and statistics team includes some of the world's most distinguished scientists and internationally recognized leaders in diverse areas such as test validity, language learning and testing, and educational measurement and statistics.

To date, more than 150 peer-reviewed TOEFL research reports, technical reports and monographs have been published by ETS, many of which have also appeared in academic journals and book volumes. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), comprised of language learning and testing experts from the academic community, funds an annual program of TOEFL research by external researchers from all over the world, including preeminent researchers from Australia, the UK, the US, Canada and Japan.

In Series One of the TOEFL iBT Research Insight Series, we provide a comprehensive account of the essential concepts, procedures and research results that assure the quality of scores on the TOEFL iBT test. The six issues in this Series will cover the following topics:

Issue 1: TOEFL iBT Test Framework and Development

The TOEFL iBT test is described along with the processes used to develop test questions and forms. These processes include rigorous review of test materials, with special attention to fairness concerns. Item pretesting, try outs and scoring procedures are also detailed.

Issue 2: TOEFL Research

The TOEFL Program has supported rigorous research to maintain and improve test quality. Over 150 reports and monographs are catalogued on the TOEFL website. A brief overview of some recent research on fairness and automated scoring is presented here.

Issue 3: Reliability and Comparability of Test Scores

Given that hundreds of thousands of test takers take the TOEFL iBT test each year, many different test forms are developed and administered. Procedures to achieve score comparability on different forms are described in this section.

Issue 4: Validity Evidence Supporting Test Score Interpretation and Use

The many types of evidence supporting the proposed interpretation and use of test scores as a measure of English-language proficiency in academic contexts are discussed.

Issue 5: Information for Score Users, Teachers and Learners

Materials and guidelines are available to aid in the interpretation and appropriate use of test scores, as well as resources for teachers and learners that support English-language instruction and test preparation.

Issue 6: TOEFL Program History

A brief overview of the history and governance of the TOEFL Program is presented. The evolution of the TOEFL test constructs and contents from 1964 to the present is summarized.

Future series will feature summaries of recent studies on topics of interest to our score users, such as "what TOEFL iBT test scores tell us about how examinees perform in academic settings," and "how score users perceive and use TOEFL iBT test scores."

The close collaboration with TOEFL iBT score users, English language learning and teaching experts and university professors in the redesign of the TOEFL iBT test has contributed to its great success. Therefore, through this publication, we hope to foster an ever stronger connection with our score users by sharing the rigorous measurement and research base and solid test development that continues to ensure the quality of TOEFL iBT scores to meet the needs of score users.

Xiaoming Xi
Senior Research Scientist
Research & Development Division
Educational Testing Service

Contributors

The primary author of this section is Mary Enright.

The following individuals also contributed to this section by providing their careful review as well as editorial suggestions (in alphabetical order).

Cris Breining
Donald Powers
Xiaofei Tang
Mikyung Kim Wolf
Brent Bridgeman
Rosalie Szabo
Eileen Tyson
Xiaoming Xi

TOEFL Research

The TOEFL® Program and the TOEFL Board have long recognized and supported the importance of research in maintaining and improving test quality. Since the mid-1970's, a portion of the annual TOEFL budget has been committed to fund and disseminate research on issues related to language assessment. ETS and the TOEFL Board support a research program to further knowledge in the field of language assessment and second-language acquisition. The goals are to:

- improve language assessments and related products
- assure that they meet professional standards
- develop the foundation for new products and services

The TOEFL Committee of Examiners (COE), a committee of the TOEFL Board composed of language research specialists from the international academic community, works closely with ETS on its program of research.

The Research Process

TOEFL research is carried out in consultation with the TOEFL COE. The COE advises the TOEFL Program about research needs and, through the Research Subcommittee, solicits, reviews and approves proposals for funding and reports for publication. The TOEFL Program also funds an extensive program of research conducted by ETS staff.

To encourage external experts to conduct TOEFL research, an announcement of the TOEFL COE research program, describing high-priority research topics, is published annually. Applications are invited from faculty or staff members who are affiliated with not-for-profit organizations and institutions (e.g., universities) with expertise in English language learning and assessment research. The COE Research Subcommittee reviews the preliminary applications. Invitations to submit a full proposal are issued to selected applicants based on the quality of the précis. Précis are evaluated in terms of the relevance to the identified topics, the feasibility and quality of the proposed research, the qualifications of the Principal Investigator, organizational capacity to conduct the research and cost effectiveness.

The quality of TOEFL research is ensured through a rigorous review process. Three to four ETS and external experts review proposals and reports. The reviewers may include applied linguists, psychologists, statisticians, psychometricians or assessment specialists. After reports are reviewed, researchers are encouraged to disseminate their findings through publications in professional journals or as TOEFL reports.

The TOEFL Program also provides a variety of other monetary grants and awards to recognize and support significant activities or projects related to the field of international education or English-language education.

Small grants are available to promising students working in the area of foreign- or second-language assessment that will help them finish their dissertations in a timely manner. Grants are also available to encourage the broad dissemination of information on English-language testing, teaching or teacher education through presentations at conferences outside the United States.

Information about TOEFL Grants and Awards is published at **www.ets.org/toefl/grants**.

Description of Selected TOEFL Research

Over 150 TOEFL research reports have been published by ETS (**www.ets.org/toefl/research**). Certain research topics such as test validation, fairness and reliability have been repeatedly re-examined over time as test methods and content evolved. Other topics represent innovations in testing such as advances in psychometrics, automated scoring and computer-based testing. And some projects have focused on the implications of theories of language proficiency for test design.

A complete listing of TOEFL reports is available. Recent reports are categorized in *Framework for Recent TOEFL Research* (ETS, 2008a), which includes links to downloadable copies of the reports. More than 50 recent reports are organized by the following categories:

- Validity Evidence
- Fairness and Accessibility
- Support for Test Revision
- Scoring and Technology
- Candidates and Populations
- Reliability and Generalizability
- Score Interpretation

A classification of earlier TOEFL reports as well as abstracts is found in *The Researcher* (ETS, 2005). Other reports on general educational and measurement topics are also published by ETS. Searches for reports on a variety of educational and measurement topics can be carried out at **http://search.ets.org/custres/**.

A comprehensive summary of all the research sponsored by ETS and the TOEFL Board is well beyond the scope of this document, and no such review will be undertaken here. A selective presentation will be made concentrating on topics not reviewed in other sections of this manual or other publications. The extensive program of research to improve language assessment that resulted in the TOEFL Internet-based Test (TOEFL iBT™) is documented in a book edited by Chapelle, Enright, and Jamieson (2008). Summaries of research and procedures to ensure that TOEFL complies with professional standards for validity (ETS, 2008b) and reliability are available (ETS, 2009). In this section we will focus on recent research concerning (a) test fairness and (b) automated analysis of writing and speaking.

Research on Test Fairness

Fairness in testing is an important measurement standard that the TOEFL Program strives to meet. For the TOEFL test, test fairness means that the test scores can be interpreted as a measure of academic English-language ability for various groups of test takers. Fairness requires that test scores should not be affected by factors that are not relevant to this intended interpretation. While care is taken during test development to ensure that test content meets fairness guidelines, empirical research studies are also conducted to determine the impact of various factors on test scores. Three recent studies addressed issues related to fairness: (a) the structure of the test for different groups of test takers, (b) the impact of educational and cultural background on reading performance, and (c) the performance of native English-speaking college students on the TOEFL iBT test.

One fairness issue concerns the factor structures of test scores for different groups. Exploratory or confirmatory factor analyses can be used to determine the underlying structure of scores on a test. The factor structure of a test should be consistent with the theoretical structure implied by the test construct. It also has implications for how scores should be reported and interpreted. Stricker and Rock (2008) analyzed the factor structure of a 2003-2004 TOEFL iBT field test form for three groups. Test takers were grouped according to (a) whether their first language was from an Indo-European vs. a non-Indo European family, (b) how widely English was used in education and business contexts in their native countries, and (c) years of studying the English language in school. The same test structure was found for all subgroups. This structure consisted of four first order factors corresponding to the four test sections (Reading, Listening, Speaking and Writing) and a higher-order factor that encompassed all the first-order factors. This structure was consistent with the theoretical model underlying the test and with the policy of reporting four section scores as well as a single composite score. The evidence of the invariance of factor structure across groups indicated that the test measures the same constructs for the groups studied and that score aggregation and reporting procedures lead to appropriate score interpretations for these groups.

Another issue with respect to the fairness of the TOEFL iBT test is whether factors other than English-language proficiency impact test performance. Liu, Schedl, Malloy and Kong (2009) viewed this as a particular concern for the TOEFL iBT Reading section, which had fewer but longer reading passages than previous versions of the TOEFL test. Their concern was that the decreased topic variety might increase the likelihood that test takers' familiarity with the particular content of a given passage would influence their reading performance on the test. Accordingly, they investigated whether TOEFL iBT reading performance was

affected by test takers' outside knowledge, gained either through academic major or from immersion in a particular culture. Performance on six passages and associated questions from five TOEFL iBT test administrations were examined. Three of the passages focused on topics in physical science, and the rest emphasized European or Japanese cultures. Differential item functioning (DIF) and differential bundle functioning (DBF) were used to investigate the impact of outside knowledge on TOEFL iBT reading performance. DIF occurs for an item when differences in performance exist after examinees are matched on the abilities that the item is intended to measure. Liu et al. found little evidence that the sources of outside knowledge they investigated influenced performance overall on the reading passages. Further, the analysis of the items displaying DIF suggests that the differences in performance may be construct-relevant differences that TOEFL iBT is intended to measure (e.g., vocabulary knowledge). To ensure continued fairness, the researchers made some recommendations to carefully scrutinize passages with technical vocabulary or culture-specific knowledge in the future.

A third fairness concern is that the TOEFL iBT test, with its academic content and tasks that required integrating different language skills, might be very difficult even for native English speakers. Native speakers, overall, do not represent the "ultimate criterion group for an ESL test, because they vary in formal and informal education in English and in linguistic ability..." (Stricker, 2002, p. 1). Nevertheless, if educated native English speakers cannot do as well as educated non-native speakers on the TOEFL iBT test, it might be claimed that non-native speakers are being held unfairly to a higher standard in admissions decisions than native speakers. In a recent study, Cline and Powers (2009) compared the performance of first-year college students who were native speakers of English with that of non-native speakers. They administered one form of the 2003-2004 TOEFL iBT field test to more than 900 first-year, native English-speaking students at community colleges and non-selective four-year colleges and compared their performance with that of the non-native speakers who had completed the field study form. Overall, the native English-speaking college students performed better than non-native speakers although there was a reasonable amount of variation in scores within this group. The mean score differences favoring the native English speakers were moderate for listening, reading and writing while they were large for speaking and for the total score. The implications are that the TOEFL iBT is neither inappropriately difficult for non-native English speakers, nor is it inordinarily easy for native English speakers. This suggests that non-native speakers are being held to a high standard, but not an unfair one.

In sum, these three studies of test structure, test content and native-speaker performance illustrate some of the fairness issues that have been addressed empirically through TOEFL research.

Automated Scoring for Writing and Speaking

There are two needs that arise when a test includes many extended constructed-response tasks such as the writing and speaking tasks on the TOEFL iBT test. One of these is the need to score the responses efficiently and reliably. The other is to provide test takers with opportunities to practice and receive feedback on their performance prior to taking the test. ETS and the TOEFL Program have been laying the foundation for new products and services that address these needs through research on automated scoring of writing and speaking. Capabilities developed at ETS that address these needs include the *e-rater*® engine, automated scoring engine and SpeechRater℠ engine.

e-rater® engine

e-rater engine uses natural language processing methods to provide feedback on the quality of students' writing and automated scores on their essays. *e-rater* engine includes a set of writing analysis tools that identify errors in grammar, usage and mechanics, as well as an essay's discourse structure and undesirable stylistic features. *e-rater* engine uses these features along with measures of the vocabulary used in the essay to statistically model human holistic ratings and provide scores on essays. These capabilities are used in combination either to rate test takers' essays on large-scale standardized tests or in practice and learning products such as ***Criterion*® *Online Writing Evaluation Service*** and **TOEFL® Practice Online**. The ***Criterion*** service is a web-based instructional tool that helps students plan, write and revise essays, and provides instant scoring and annotated diagnostic feedback. TOEFL Practice Online is

a practice test for TOEFL iBT that provides students with instant scores and performance feedback.

An extensive program of research, documented at the ***Criterion*** service and the ***e-rater*** engine, contributed to the continuous development and refinement of these capabilities and their evaluation for use in different contexts. While this research initially focused on analyzing and scoring essays written primarily by native English speakers (e.g., Kaplan R. A., Wolff, Burstein, Lu, Rock, & Kaplan, B.A., 1998), attention soon expanded to include research on essays written specifically by non-native English speakers (e.g., Chodorow & Burstein, 2004). During the past decade, such research has addressed two questions.

The first question is whether the use of the *e-rater* engine in conjunction with human ratings to score the TOEFL iBT writing tasks is justified. In their summary of research relevant to the use of the *e-rater* engine for the independent writing task, Enright and Quinlan (2010) report that the *e-rater* engine has been found to agree with human raters as well as or better than human raters agree with each other when rating these essays. They also address the issue of the correspondence between the qualities and processes used by humans to rate these essays, and the features and processes used by the *e-rater* engine. They conclude that humans and the *e-rater* engine have complementary strengths. When humans and the *e-rater* engine are compared, the *e-rater* engine assesses a more limited range of essay qualities than do humans but it does so in a more consistent manner. Finally, there is some evidence that ratings by the *e-rater* engine and by humans have similar relationships with other criteria of language ability, especially those that reflect writing ability. Overall, the empirical evidence summarized by Enright and Quinlan supports the use of the *e-rater* engine as a complement to human raters to score TOEFL independent essays. Research has also been conducted to evaluate the use of the *e-rater* engine for the integrated writing task, which requires test takers to summarize and synthesize academic reading and listening materials in writing. The areas of research included the degree of agreement of the *e-rater* engine with human scores, the relationships of human and the *e-rater* engine scores to independent indicators of language ability, and the impact of using the *e-rater* engine on scores by demographic subgroup. The results yielded supportive evidence to use the *e-rater* engine to complement human raters for TOEFL integrated writing task as well. These studies will be available on the ETS website in the near future. Further research is ongoing to enhance and expand the the *e-rater* engine capabilities for analyzing the content, organization and coherence of essays.

Other researchers have explored the second question of whether the *e-rater* engine has the potential to provide analytic trait scoring of essays from the TOEFL test. Analytic trait scoring, which provides measures of independent writing subskills, appeals to writing teachers as a guide to instruction. However, traits, when scored by humans, are often found to be highly correlated and therefore indistinguishable. Because the *e-rater* engine feature values in scoring models are less correlated, they have the potential to provide more independent trait scores. Attali (2007) conducted a factor analysis of the *e-rater* engine features in a scoring model used to rate essays from the TOEFL test. He found three factors corresponding to discourse, word usage and grammar. Attali suggested that, given the low correlations among these factors, they have the potential to describe three independent traits. Lee, Gentile and Kantor (2010) investigated the relationship between human analytic trait scores and the *e-rater* engine feature scores. They concluded *"To a certain degree, it seems also justifiable to use some of these existing e-rater variables to compute automated trait scores representing different aspects of essay quality"* p. 21. In light of these results, further investigation of the e-rater's engine 's potential to provide analytic trait scoring in the context of large-scale standardized assessment is a promising direction for future research.

SpeechRater℠ System

Automated scoring of speech is a more recent development than automated scoring of writing and presents a greater challenge because of the difficulty of automatically recognizing the words in continuous speech. While speech scoring systems for simple tasks that require the production of a limited or predictable range of vocabulary have been in use for a number of years (see Zechner, Higgins, Xi & Williamson, 2009 for review), the tasks on the TOEFL Speaking section are more complex. The Speaking section includes six tasks that require test takers either to respond to a relatively general question or to respond to oral and/or written input. While TOEFL iBT spoken responses are scored holistically by raters using a four-point scale, the raters are instructed to attend to three key aspects of performance:

delivery, language use and topic development (see TOEFL iBT Test Framework and Test Development). Given the complexity of the TOEFL speaking tasks and scoring guidelines, a number of factors make the automated analysis of the TOEFL speaking samples difficult. These are (a) the length of the responses (45 to 60 seconds), (b) variability in topics, discourse structure and lexical choices, (c) great diversity in test takers' accents, (d) the wide range of proficiencies in speaking and (e) the different aspects of response to be considered in scoring.

The *SpeechRater* system, developed at ETS, is used currently to automatically score responses to TOEFL iBT speaking tasks in a practice environment (Zechner et al., 2009). The system consists of three components: a speech recognizer, a feature computation model and a scoring model. The speech recognizer was trained on responses by non-native English speakers to TOEFL iBT speaking tasks in a practice context. The feature computation model uses the output of the speech recognizer to compute a set of features. And the scoring model uses these features to predict statistically a score for each response.

The developmental research supporting the *SpeechRater* system has addressed many aspects of system quality, including the construct coverage of the scoring features and the prediction accuracy of the scoring model (Zechner, Higgins & Xi, 2007; Zechner et al., 2009). The speech recognizer provides information about word identity and timing. The system developers and language experts identified a set of 29 construct relevant features that could be extracted from the output of the speech recognizer. These features were consistent with the construct of communicative language ability as embodied in the scoring guidelines. Features identified were related primarily to the delivery aspect of the guidelines and focused on fluency and pronunciation. A few features were related to language use, including vocabulary diversity and some aspects of grammar. However, because speech recognizer accuracy in word identification is only moderate (50%), features did not cover topic development. This set of features was further reduced by considering the conceptual overlap and the intercorrelations among the features and their empirical relationship to human scores. A final step was to develop a statistical model that not only predicted human scores but also provided reasonable coverage of the communicative construct embodied in the scoring guidelines. This initial version of the *SpeechRater* system predicts human scores well enough for use in a practice environment. For data from the TOEFL Practice Online Speaking section, the correlation between the *SpeechRater* scores and human scores was .57, while that between two human raters was .74. The relatively moderate correlations were partially due to the limited variability in the scores in the TOEFL Practice Online data (on a score scale of 1-4, most scores were 2 and 3). For a dataset that had more variability in the scores (i.e., the TOEFL Field Study data), the human-*SpeechRater* system score correlation increased to .68.

Developmental research on the *SpeechRater* system is ongoing. Goals are to improve the accuracy of the speech recognizer, develop features to provide better coverage of the construct and to improve the agreement of the SpeechRater scores with those of human raters.

This brief description of a few studies does little to convey the extent of the contribution that ETS and the TOEFL Program have made to advancing knowledge of language assessment. The descriptions of more than 150 research studies available through the TOEFL website illustrate the Program's commitment to advancing the field and meeting high standards for educational measurement.

References

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

Chapelle, C. A., Jamieson, J., & Enright, M. K. (Eds.) (2008). Building a validity argument for the Test of English as a Foreign Language. London: Routledge.

Cline, F. & Powers, D. E. (2009). The new generation TOEFL: Evaluating its use with native speakers of English. Manuscript submitted for publication.

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

Enright, M. K. & Quinlan, T. (2010). Complementing human judgment of essays written by English Language Learners with E-rater® scoring. Language Testing.

ETS (2005). The Researcher. Retrieved February 16, 2010, from **http://www.ets.org/Media/Research/pdf/TheResearcher2005.pdf**

ETS. (2008a). Framework for Recent TOEFL Research. Retrieved February 16, 2010, from **http://www.ets.org/Media/Research/pdf/Framework_Recent_TOEFL_Research.pdf**

ETS. (2008b). Validity evidence supporting the interpretation and use of TOEFL® iBT scores. Retrieved February 16, 2010, from **http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Validity.pdf**

ETS. (2009). Reliability and comparability of TOEFL® iBT scores. Retrieved February 16, 2010, from **http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Reliability.pdf**

Kaplan, R. M., Wolff, S.E., Burstein, J. C., Lu, C., Rock, D.A., & Kaplan, B.A. (1998). Scoring essays automatically using surface features (GRE Board Professional Rep. No 94-21P; ETS RR-98-39). Princeton, NJ: ETS.

Lee, Y.-W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31 (3), 391-417.

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). Does content knowledge affect TOEFL iBT™ reading performance? A confirmatory approach to differential item functioning. TOEFL iBT Research Report No. 09. Princeton, NJ: ETS.

Stricker, L. J. (2002). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. TOEFL Research Report No. 69. Princeton, NJ: ETS.

Contact Us toeflnews@ets.org

TOEFL iBT™ Research • Series 1, Volume 2

Insight

760110

Listening. Learning. Leading.®

www.ets.org