

TOEFL iBT™ Research

Insight

Series I, Volume 4

Validity Evidence Supporting the
Interpretation and Use
of TOEFL iBT™ Scores



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 authors of this section are Mary Enright and Eileen Tyson.

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

Cris Breining
Rosalie Szabo
Xiaofei Tang
Mikyung Kim Wolf
Xiaoming Xi

Validity Evidence Supporting the Interpretation and Use of TOEFL iBT™ Scores

Validity is “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999, p. 9).

Test validation is the process of making a case for the proposed interpretation and uses of test scores. This case takes the form of an argument that states a series of propositions supporting the proposed interpretation and uses of test scores, and summarizes the evidence supporting these propositions (Kane, 2006). Because many types of evidence are relevant to these propositions, validation

For the TOEFL iBT™ test, the validation process began with the conceptualization and design of the test

requires an extended research program. For the TOEFL iBT™ test, the validation process began with the conceptualization and design of the test (Chapelle, Enright, & Jamieson, 2008) and continues today with an ongoing program of validation research as the test is being used to make decisions about test takers’ academic English language proficiency.

TOEFL iBT test scores are interpreted as the ability of the test taker to use and understand English as it is spoken, written, and heard in college and university settings. The proposed uses of TOEFL iBT test scores are to aid in admissions and

placement decisions at English-medium institutions of higher education and to support English-language instruction.

In this document we lay out the validity argument for the TOEFL iBT test by stating the propositions that underlie the proposed test score interpretation and uses, and by summarizing the evidence relevant to each proposition (see Table 1).

Table 1.
Propositions and Related Evidence in the TOEFL Validity Argument

Proposition ¹	Evidence
The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.	Reviews of research and empirical studies of language use at English-medium institutions of higher education
Tasks and scoring criteria are appropriate for obtaining evidence of test takers’ academic language abilities.	Pilot and field studies of task and test design; systematic development of rubrics for scoring written and spoken responses
Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks.	Investigations of discourse characteristics of written and spoken responses and strategies used in answering reading comprehension questions
The structure of the test is consistent with theoretical views of the relationships among English language skills.	Factor analyses of a field-study test form
Performance on the test is related to other indicators or criteria of academic language proficiency.	Relationships between test scores and self-assessments, academic placements, local assessments of international teaching assistants, performance on simulated academic listening tasks
The test results are used appropriately and have positive consequences.	Development of materials to help test users prepare for the test and interpret test scores appropriately; long-term empirical study of test impact (washback)

¹ Another important proposition in the TOEFL validity argument, that test scores are reliable and comparable across test forms, is the subject of a separate article (ETS, 2008).

In the following sections we describe the evidence relevant to these propositions. The collection of this evidence began with the initial discussions about a new TOEFL test in the early 1990's and continues today through an active research program. The initial discussions about a new TOEFL test led to many empirical investigations and evaluations of the results and the emerging test design by experts in language testing. Prototyping, usability, and pilot studies were conducted from 1999 to 2001. Two large-scale field studies were carried out in the spring of 2002 and the winter of 2003-2004. Validity research continues today now that the TOEFL iBT test is operational.

The Relevance and Representativeness of Test Content

The first proposition in the TOEFL validity argument is that the test content is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings. Because the primary use of TOEFL test scores is for admissions to English-medium colleges and universities, score users often want evidence that supports this proposition. Score users want evidence that the test content is “authentic.”

Nevertheless, tests are events that are distinct from other academic activities. Test tasks and content are likely to be simulations, but not exact replications, of academic tasks. Accordingly, the test design process began with the analysis of the academic tasks and identification of important characteristics of these tasks that could be captured in test tasks. This analysis focused on the knowledge, abilities, and skills needed to succeed in academic situations as well as the tasks and materials typically encountered in colleges and universities. The development of the TOEFL iBT test began with reviews of research about the English-language skills needed for study at English-medium institutions of higher education. Subsequently, groups of experts laid out preliminary frameworks for a new test design and associated research agendas. This groundwork for the new test is summarized by Taylor and Angelis (2008) and Jamieson, Eignor, Grabe, & Kunnan (2008).

Research that supported the development of relevant and representative test content included three empirical studies:

Rosenfeld, Leung, & Oltman (2001) helped establish the importance of a variety of English-language skills and tasks for academic success through a survey of undergraduate and graduate faculty and students. These data on faculty and student judgments of the relative importance of tasks were taken into consideration in the design of test tasks and criterion measures.

Biber and his associates (Biber et al., 2004) helped establish the representativeness and authenticity of the lectures and conversations that are used to assess listening comprehension on the TOEFL iBT test. But this work also demonstrated constraints on the degree of authenticity that can characterize test tasks. Biber et al. collected a corpus of 1.67 million words of spoken language at four universities. The linguistic features of this corpus were then analyzed to provide guidelines for the characteristics of the lectures and conversations to be used on the TOEFL iBT test. However, unedited excerpts of authentic samples of aural language from the corpus could not be used for test materials for a number of reasons. Many excerpts required students to have knowledge other than that of the English language (e.g., mathematics, discipline-specific knowledge, information presented in other related sources) or contained references to American culture that might not be understood internationally, or presented topics that might be upsetting to some students.

Cumming, Grant, Mulcahy-Ernt, & Powers (2005) provided evidence about the content relevance, authenticity, and educational appropriateness of “integrated” test tasks. One of the most innovative aspects of the TOEFL iBT test was the introduction of tasks that required the integrated application of two or more language skills. On the speaking and writing sections of the test, some tasks require test takers to incorporate information from a brief lecture and short reading passage into their spoken or written responses. As preliminary versions of these integrated tasks were considered for inclusion on the test, Cumming et al. interviewed a sample of English as a second language (ESL) teachers about these new types of test tasks. The teachers viewed the tasks positively, judging the tasks to be realistic and appropriate simulations of academic tasks. They also felt that the tasks elicited speaking and writing samples from their students that represented the way the students usually performed in their English classes.

These teachers' comments about how the tasks could be improved also influenced further refinements of the task characteristics.

Task Design and Scoring Rubrics

The design and presentation of tasks, and the rubrics (criteria) used to score responses, need to be appropriate for providing evidence of test takers' academic language abilities. The developers of the TOEFL iBT test carried out numerous exploratory studies over four years to determine the best way to design, administer, and score new assessment tasks (Chapelle, Enright, & Jamieson, 2008). The initial studies about task design and presentation informed decisions about:

- The characteristics of the reading passages and listening materials
- The types of tasks used to assess reading and listening
- The types of integrated tasks used to assess speaking and writing
- The computer interface used to present the tasks
- The use of note taking
- The timing of the tasks
- The number of tasks to include in each section

Careful attention was paid to the development of rubrics (criteria) to score the responses to speaking and writing tasks. Groups of experts reviewed task takers' responses and proposed scoring criteria. Investigations of raters' cognitive processes as they analyzed test takers' responses also contributed to the development of these scoring rubrics (Brown, Iwashita, & McNamara, 2005; Cumming et al., 2006). The rubrics were then trialed in field studies and revised to result in four-point² holistic rubrics for speaking (ETS, 2004a), and five-point² holistic rubrics for writing (ETS, 2004b).

² In addition, irrelevant or absent responses receive a score of 0.

Linguistic Knowledge, Processes, and Strategies

Another proposition, that academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks, is supported by three studies to date. These studies included investigations of the discourse characteristics of test takers' written responses, of their spoken responses, and of verbal reports by test takers as they responded to reading comprehension questions.

For writing and speaking tasks, the characteristics of the discourse that test takers produce is expected to vary with score level as described in the holistic rubrics that raters use to score responses. Furthermore, the rationale for including both independent and integrated tasks in the TOEFL iBT speaking and writing sections was that these types of tasks would differ in the nature of discourse produced, thereby broadening representation of the domain of academic language on the test.

Cumming et al. (2006) analyzed the discourse characteristics of a sample of 36 examinees' written responses to prototype independent and integrated essay questions. For independent tasks, writers were asked to present an extended argument drawing on their own knowledge and experience. For integrated tasks, writers were asked to respond to a question drawing on information presented in a brief lecture or reading passage. Cumming found that the discourse characteristics varied as expected, both with writers' proficiency levels and with task types. Discourse features analyzed included text length, lexical sophistication, syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source text. Greater writing proficiency (as reflected in the holistic scores previously assigned by raters) was associated with longer responses, greater lexical sophistication, syntactic complexity, and grammatical accuracy. In contrast with the independent tasks, responses to the integrated tasks had greater lexical sophistication and syntactic complexity, relied more on the source materials for information, and used more paraphrasing and summarization.

Discourse analyses of responses to early prototypes of independent and integrated speaking tasks were also carried out (Brown, Iwashita, & McNamara, 2005). The prototype tasks included two independent tasks and three integrated ones. The latter tasks drew on information presented in either a lecture or a reading passage. Two hundred speech samples (40 per task), representing five proficiency levels, were analyzed. Speech samples were coded for discourse features representative of four major conceptual categories: linguistic resources, phonology, fluency, and content. Brown et al. found that the qualities of spoken responses varied modestly with proficiency level, and a lesser amount with task type. Greater fluency, more sophisticated vocabulary, better pronunciation, greater grammatical accuracy, and more relevant content were characteristics of speech samples receiving higher holistic scores from raters. When compared with responses to independent tasks, responses to integrated tasks had a more complex schematic structure, were less fluent, and included more sophisticated vocabulary.

An investigation of strategies used by test takers to answer reading comprehension questions was carried out by Cohen and Upton (2006). Verbal report data were collected from 32 students, representing four language groups (Chinese, Japanese, Korean, and Other languages), as they responded to prototype TOEFL reading comprehension tasks closely resembling tasks now on the TOEFL iBT test. The reading and test-taking strategies evident in the students' verbal reports were analyzed. In summarizing the reading and test-taking strategies that were used for the full range of questions types, the authors noted that test takers did not rely on "test wiseness" strategies. Rather their strategies,

reflect the fact that respondents were in actuality engaged with the reading test tasks in the manner desired by the test designers...respondents were actively working to understand the text, to understand the expectations of the questions, to understand the meaning and implications of the different options in light of the text, and to select and discard options based on what they understood about the text. (p. 105)

These findings help to refute a concern that test takers might receive high scores on reading comprehension tests primarily by using "test wiseness" strategies (e.g., matching of words in the question to the passage without understanding) rather than reading strategies (e.g., reading the passage carefully) or appropriate test management strategies (e.g., selecting options based on meaning).

Test Structure

Factor analytic studies provide evidence that the structure of the test is consistent with theoretical views of the relationships among English-language skills. The TOEFL iBT test is intended to measure a complex, multi-componential construct of English as a foreign language (EFL) ability, consisting of a general English-language ability factor as well as other factors associated with specific language skills. Validation research as to whether the test actually measures the intended model of the construct was conducted with confirmatory factor analysis of a 2003-2004 TOEFL iBT field study test form (Sawaki, Stricker, & Oranje, 2008). The researchers reported that the factor structure of the test was best represented by a higher-order factor model with a general factor (EFL ability) and four group factors, one each for reading, listening, speaking, and writing. These empirical results are consistent with the intended model of English-language abilities. That is, there are some aspects of English-language ability common to the four skills, as well as some aspects that are unique to each skill. This finding is consistent with the way test scores are reported and used. This higher-order factor structure also proved to be invariant across subgroups who took this test form and who differed by (a) whether their first-language background was Indo-European or Non-Indo-European and (b) their amount of exposure to English (Stricker & Rock, 2008).

Relationship between TOEFL iBT Scores and Other Criteria of Language Proficiency

Another important proposition underlying valid score interpretation and use is that the performance on the test is related to other indicators or criteria of academic language proficiency.

A central question for test users interpreting test scores is, “Does a test score really

A central question for test users interpreting test scores is, “Does a test score really tell me about a student’s performance beyond the test situation?”

tell me about a student’s performance beyond the test situation?” “Is a student just a good test taker when it comes to TOEFL iBT? Or do TOEFL scores really indicate whether or not the student has a level of English-language proficiency sufficient for study at an English-medium college or university?”

The answer lies in evidence demonstrating a relationship between test scores and other measures or criteria of language proficiency. One challenge, of course, is to determine what these “other” criteria should be. For many admission tests for higher education, which are intended to assess broader academic skills and to predict success in further studies, the grade point average in college or graduate school often serves as a relevant criterion. However, the TOEFL test is intended to measure a narrower construct of academic *English-language* proficiency. Therefore, grades averaged across all academic subjects would not be appropriate as a criterion for the TOEFL iBT test, particularly grades from different education systems around the world.

A second issue concerns the magnitude of statistical effects: How strong a relationship between test scores and other criteria should be expected? Correlations are the statistic most often used to describe the relationship between test scores and other criteria of proficiency. But two factors constrain the magnitude of such correlations. One is that criterion measures often have low reliability, limiting the degree of correlation they can have to test scores. Another is method effects: The greater the difference between the kinds of measures being compared (e.g., test scores versus grades in courses), the lower the correlations will be. For instance, a test may assess a relatively specific

academic skill, whereas grades in courses may be affected by a broader range of students’ characteristics, such as study skills and motivation. As an example, correlations between similar types of measures, TOEFL CBT and iBT, are quite high (observed $r = .89$, Wang, Eignor, & Enright, 2008). However, the typical correlation between different types of measures, such as aptitude test scores and school grades, are more modest, on the order of $r = .5$ (Cohen, 1988).

As the TOEFL iBT test was being developed, relationships between test scores and other relevant criteria of academic language proficiency were investigated. These other criteria included the following:

Self assessment. The participants in the 2003-2004 field study of a TOEFL iBT test form were asked to indicate how well they agreed with a series of “can do” statements on a questionnaire (Wang, Eignor, & Enright, 2008). These statements represented a range of complexity in language tasks. As an example, a statement about a simple task for speaking was, “My instructor understands me when I ask a question in English.” A statement about a more complex speaking task was “I can talk about facts or theories I know well and explain them in English.” There were 14 to 16 such statements for each of the four language skills (listening, reading, speaking, and writing), and over 2,000 test takers completed the questionnaire. Observed correlations between the summative scores for each of the four self-assessment scales averaged .46 with test scores on the measures of four skills, and .52 with the total test score. Moreover, test takers with higher test scores were more likely to indicate that they could do more complex tasks than were test takers with lower test scores (ETS, 2004c).

Academic placement. The relationship between TOEFL iBT scores and academic placement at colleges and universities also provides evidence that the test scores are related to other indicators of academic language proficiency. In many English-medium colleges and universities, some international students are judged to have sufficient English-language skills to take content courses without needing additional English-language instruction. Other international students, who are judged to be less proficient in English, are required to take English as a second language (ESL) development courses in addition to their content courses. Still other students themselves may enroll in intensive English programs (IEPs), hoping to improve their English-language skills to prepare themselves for university study.

These placements into ESL language development courses and IEPs reflect a lower level of English-language proficiency than does unrestricted enrollment in content courses. In the 2003-2004 field study, test takers who were studying in English-speaking countries were asked about their academic placement (Wang, Eignor, & Enright, 2008). Differences in test scores between students who were enrolled in ESL language development courses or IEPs, and those enrolled in only content courses, were large and statistically significant, as illustrated in Figure 1.

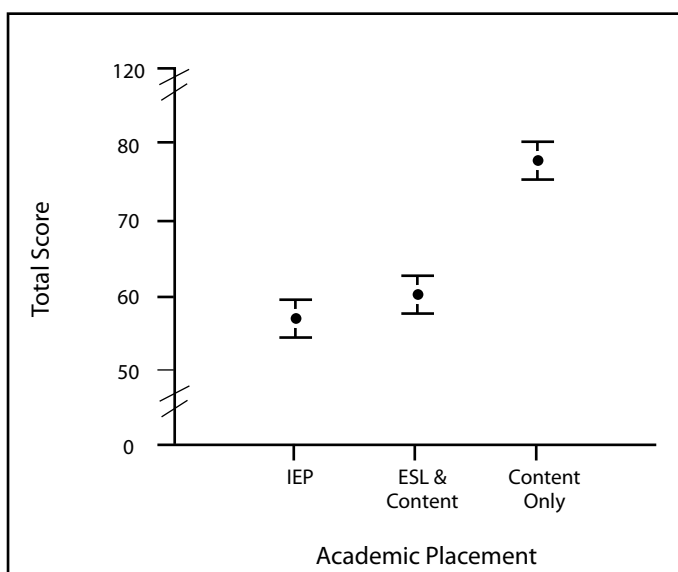


Figure 1. Mean total test score (+/- 95% CI) for test takers studying in English-speaking countries who were enrolled in institutional English programs (IEP, $n = 208$), taking both ESL courses and content courses ($n = 285$) or only content courses ($n = 488$).

Local Institutional Tests for International Teaching Assistants.

The most common use of TOEFL iBT scores is to aid in the admissions process, but the speaking score is potentially useful as a pre-arrival screening measure for international teaching assistants (ITAs). Xi (2008) investigated whether or not scores on the TOEFL iBT speaking section could help institutions distinguish between candidates whose English was or was not sufficient to begin teaching. Xi examined the relationship between scores on the TOEFL iBT speaking section and on local tests used for this purpose after candidates arrived at their universities. One characteristic of the local tests was that they used performance-based assessments that attempt to simulate English language use in instructional settings. The observed correlations between scores on the TOEFL iBT speaking section and on these local ITA assessments are presented in Table 2.

Table 2.
Correlations between the Scores on the TOEFL iBT Speaking Section and Different Types of Local ITA Assessments

Type of Local ITA Assessment	Observed Correlation
<i>Simulated teaching test (content and non-content combined) scored on the basis of linguistic qualities (n = 84)</i>	.78
<i>Simulated teaching test (separate content and non-content based tests) scored on the basis of linguistic qualities and teaching skills (n = 45)</i>	.70
<i>Simulated teaching test (content-based) scored on the basis of linguistic qualities, teacher presence, and nonverbal communication (n = 53)</i>	.53
<i>Real classroom teaching sessions scored on the basis of linguistic qualities, lecturing skills, and awareness of American classroom culture (n = 23)</i>	.44

Xi noted that the strength of the relationship was affected by the extent to which the local ITA tests engaged and evaluated non-language abilities. The more the assessment focused on speaking abilities and the less on teaching skills, the higher was the correlation between the scores on local ITA assessments and the TOEFL iBT speaking section. This is consistent with the intended interpretation of the TOEFL speaking score as a measure of language ability, not teaching ability.

Performance on Simulated Academic Listening Tasks.

Given the challenge of finding appropriate, existing criteria of academic listening ability, Sawaki and Nissan (2009) created their own criterion, a set of three complex academic

listening tasks. They found that the TOEFL iBT listening score corresponded well with performance on these tasks. Sawaki and Nissan carefully constructed simulated academic listening tasks by surveying a sample of undergraduate and graduate students at four universities about the importance of a variety of academic listening tasks

The aim of the TOEFL iBT test is to maximize the positive consequences of score use.

and course-related activities and assignments. This survey indicated that the most frequent and important listening activity was listening to instructors presenting academic materials. Answering objective and short-answer questions was the most frequent class assignment and the most important component of final grades.

Based on this survey, Sawaki and Nissan constructed three simulated academic tasks that each consisted of a 30-minute lecture followed by a set of listening comprehension questions. The lectures were commercially available, video-based academic lectures covering introductory topics in history, psychology, and physics. The listening comprehension sets, including a total of 32 objective and short-answer questions with a maximum possible score of 44 points, were developed by content area experts in collaboration with the researchers. The scoring criteria for the short answers were also developed by the content experts. A sample of 120 graduate students and 64 undergraduates completed a TOEFL iBT listening section and the three academic listening tasks. The observed correlations between TOEFL iBT listening section scores and the simulated academic listening tasks were .56 for undergraduate students and .64 for graduate students.

Test Use and Consequences

The final proposition in the TOEFL validity argument is that the test is used appropriately and has positive consequences. In recent years, analyzing the consequences of test use has become an important aspect of test validation, and these consequences can be positive or negative. The aim of the TOEFL iBT test is to maximize the positive consequences of score use.

The primary use of the TOEFL test is to make decisions about students' readiness to study at English-medium educational institutions. From this perspective, positive consequences

would involve granting admission to students who have the English language proficiency necessary to succeed at the institution, and denying admission to those who do not. Negative consequences would involve granting admission

to students who do not have the English language proficiency necessary to succeed at the institution, and denying admission to those who do. These latter consequences are viewed as negative because, on the one hand, they waste institutional and student resources and misinform expectations, or, on the other hand, they deny opportunity to qualified students. Using test scores appropriately to make decisions with positive consequences is the joint responsibility of the test user and

ETS has been proactive in encouraging positive washback from the TOEFL iBT test on English teaching and learning.

the test publisher. To support appropriate use of TOEFL iBT scores, ETS has provided

score users with descriptive information to help them to interpret test scores (ETS, 2004c), guidance on how to set standards for using scores at their institution for admissions purposes (ETS, 2005), and an empirical study, described above, on the effectiveness of speaking scores in making decisions about ITAs (Xi, 2008).

Another intended use of the TOEFL iBT test is to support appropriate methods for teaching and learning English. One consequence of test use that has been of particular concern in the English-language teaching community has been the perceived negative impact of tests, often referred to as negative washback, on teaching and learning. Innovations in the TOEFL iBT test, such as the introduction of a speaking section and the inclusion of integrated tasks, were motivated by a belief that these innovations would prompt the creation and use of test preparation materials and activities that would more closely resemble communicatively-oriented pedagogy in academic English courses.

To this end, ETS has been proactive in encouraging positive washback from the TOEFL iBT test on English teaching and learning. A manual, *Helping Your Students Communicate with Confidence* (ETS, 2004d), was prepared for curriculum coordinators, academic directors, and teachers. The manual describes the relationship between communicative approaches to teaching English and the design of the TOEFL iBT test. It also provides sample tasks and suggestions for classroom activities. Information about the concepts underlying the test and sample materials have

been shared with textbook publishers with the intent of positively affecting the materials they produce for English-language learners.

Given that the impact of the TOEFL iBT test on teaching and learning will only be evident over time, long term research to track TOEFL-related changes is underway. Wall and Horák (2006, 2008a) have been studying how a small number of English-language teachers in Eastern Europe cope with changes in the test, and whether and how test preparation materials change in response to the new test. The investigation involves four phases:

- **Phase 1** (Wall and Horák, 2006) constituted a baseline study in which observations were carried out and interviews conducted with teachers, students, and directors at 10 institutions in six countries in Central and Eastern Europe prior to the introduction of the TOEFL iBT test. Teachers' instructional techniques were found to be highly dependent on test preparation course books that emphasized practicing the types of test items typical of the paper-based and computer-based versions of the TOEFL test. Overall the teachers were aware of the subskills that contributed to reading development, but they lacked techniques for breaking the listening down into subskills to facilitate development. Teachers devoted considerable time to teaching writing, but not speaking, as it was not viewed as an important skill to practice or learn because it was not on the test.

- **Phase 2** (Wall and Horák, 2008a) monitored six teachers from five of these countries to explore their awareness of the new TOEFL test, the features of their test preparation classes, their reactions to the most innovative parts of the new test, and their thoughts about the type of content and activities they would offer once the TOEFL iBT test was operational in their countries. The teachers' reactions to the new test were mostly positive, especially as to the idea of testing speaking. The integrated writing task was also received favorably, as was the idea that students would be able to take notes during the listening section and not have to rely on their memory. The teachers felt that these innovations would lead to changes in their classes, but most of them could only envisage changes in general terms and were waiting for test preparation materials to appear that would help them to decide on the details.

- **Phase 3** (Wall and Horák, 2008b) is focusing on an analysis of sources of information about test content and format, particularly commercial test preparation textbooks that are important mediators of test impact. The analysis will reveal the degree to which test preparation materials have been influenced by developments in the TOEFL test, and judge whether influences have been in terms of content only, or whether they provide sufficient methodological support to help teachers make teaching related to the TOEFL test more varied and stimulating than was the case in Phase 1.

- **Phase 4** (Wall and Horák, 2008b) includes observations of classroom teaching and interviews with a few of the Phase 1 teachers and directors of their institutions

Greater attention is paid to the development of speaking and there is a new focus on the integration of multiple skills in teaching.

to observe what, if any, changes in teaching have occurred. While some aspects of teaching seem not

to have changed a great deal, others have changed considerably. Greater attention is paid to the development of speaking and there is a new focus on the integration of multiple skills.

Conclusion

Although the TOEFL iBT test has only been in use since 2005, a strong case for the validity of proposed score interpretation and uses has been constructed. Concerns about test validation were an integral part of the test design process. The evidence gathered during that process has been documented and synthesized (Chapelle, Enright, & Jamieson, 2008). Even so, test validation is an ongoing process that continues to be actively supported by ETS and the TOEFL Board through the Committee of Examiners (COE) Research Program. The COE, composed of distinguished ESL experts from the academic community in North America and around the world, publishes an annual announcement of a research program and invites language teaching and testing experts to submit proposals. In this way, the case for valid score interpretation continues to grow and be refined.

References

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph No. MS-25). Princeton, NJ: Educational Testing Service.
- Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks* (TOEFL Monograph No. MS-29). Princeton, NJ: Educational Testing Service.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, A., & Upton, T. (2006). *Strategies in responding to the New TOEFL reading tasks* (TOEFL Monograph No. MS-33). Princeton, NJ: Educational Testing Service.
- Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL®* (TOEFL-MS-26). Princeton, NJ: Educational Testing Service.
- Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype writing tasks for new TOEFL* (TOEFL Monograph No. MS-30). Princeton, NJ: Educational Testing Service.
- Educational Testing Service. (2004a). *Scoring Guides (Rubrics) for Speaking*. Princeton, NJ: Author.
- Educational Testing Service. (2004b). *Scoring Guides (Rubrics) for Writing*. Princeton, NJ: Author.
- Educational Testing Service. (2004c). *English Language Competency Descriptors*. Princeton, NJ: Author.
- Educational Testing Service. (2004d). *Helping Your Students Communicate with Confidence*. Princeton, NJ: Author.
- Educational Testing Service. (2005). *Standard setting materials for the Internet-based TOEFL test*. [Compact Disk]. Princeton, NJ: Author.
- Educational Testing Service. (2008). *Reliability and Comparability of TOEFL® iBT Scores*. Princeton, NJ: Author.
- Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.) *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger.
- Rosenfeld, M., Leung, P., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. 21). Princeton, NJ: Educational Testing Service.
- Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL® iBT listening section* (TOEFL iBT Research Report No. TOEFLiBT-08). Princeton, NJ: Educational Testing Service.
- Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample*. (TOEFL iBT Research Report No. TOEFLiBT-04). Princeton, NJ: Educational Testing Service.
- Stricker, L. J., & Rock, D. A. (2008). *Factor Structure of the TOEFL Internet-based Test across Subgroups* (TOEFL iBT Research Report No. TOEFLiBT-07). Princeton, NJ: Educational Testing Service.
- Taylor, C., & Angelis, P. (2008). The evolution of TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.) *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe: Phase 1, the baseline study* (TOEFL Monograph No. MS-34). Princeton, NJ: Educational Testing Service.
- Wall, D., & Horák, T. (2008a). *The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe: Phase 2, Coping with change* (TOEFL iBT Research Report No. TOEFLiBT-05). Princeton, NJ: Educational Testing Service.
- Wall, D., & Horák, T. (2008b, April). *The TOEFL impact study*. Paper presented at the 3rd annual conference of the Association of Language Testers in Europe, Cambridge, England.
- Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.) *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs*. (TOEFL iBT Research Report No. TOEFLiBT-03). Princeton, NJ: Educational Testing Service.

Contact Us toeflnews@ets.org

TOEFL iBT™ Research • Series 1, Volume 4

Insight



Copyright © 2011 by Educational Testing Service. All rights reserved. ETS, the ETS logo, LISTENING. LEARNING. LEADING. and TOEFL are registered trademarks of Educational Testing Service (ETS) in the United States and other countries. TOEFL iBT is a trademark of ETS. ETS10101

760651



Listening. Learning. Leading.®

www.ets.org