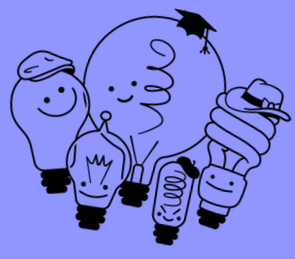
toefl®

TOEFL® Research INSIGHTS

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

VOLUME 4



TOEFL® Research Insight Series, Volume 4: Reliability and Comparability of TOEFL iBT® Scores

Preface

The TOEFL iBT® test is the world's most widely respected English language assessment, used for admissions purposes in more than 150 countries, including Australia, Canada, New Zealand, the United Kingdom, and the United States (see test review in Alderson, 2009). 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 integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments.

In addition to the TOEFL iBT test, the TOEFL® Family of Assessments was expanded to provide high-quality, English proficiency assessments for a variety of academic uses and contexts. The TOEFL® Young Students Series (YSS) features the TOEFL Primary® and TOEFL Junior® tests, which are designed to help teachers and learners of English in school settings. In addition, the TOEFL ITP® program offers colleges, universities, and others affordable tests for placement and progress monitoring within English programs as a pathway to eventual degree programs. The TOEFL Essentials test evaluates the four language skills in a friendly test format, with short, engaging tasks that relate to both academic situations and everyday life.

At ETS, we understand that scores from the TOEFL Family of Assessments are used to help make important decisions about students, and we would like to keep score users and test takers up to date about the research results that help assure the quality of these scores. Through the TOEFL® Research Insight Series, we provide institutions and English teachers with information regarding the strong research and development base that underlies the TOEFL Family of Assessments, and demonstrates our continued commitment to research.

Since the 1970s, 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, it is only through a rigorous program of research that a testing company can substantiate claims about what test takers know or can do based on their test scores, as well as provide support for the intended uses of assessments and minimize potential negative consequences of score use. Beyond demonstrating this critical evidence of test quality, research is also important for enabling innovations in test design and addressing the needs of test takers and test score users. This is why ETS has established a strong research base as a fundamental feature underlying the evolution of the TOEFL Family of Assessments.

This TOEFL Family of Assessments is designed, produced, and supported by a world-class team of test developers, educational measurement specialists, statisticians, and researchers in applied linguistics and language testing. Our test developers have advanced degrees in fields such as English, language education, and applied linguistics. They also possess extensive international experience, having taught English on continents around the globe. Our research, measurement, and statistics teams include some of the world's most distinguished scientists and internationally recognized leaders in diverse areas such as test validity, language learning and assessment, and educational measurement.

To date, more than 300 peer-reviewed TOEFL Family of Assessments research reports, technical reports, and monographs have been published by ETS, and many more studies on the TOEFL tests have appeared in academic journals and book volumes. In addition, over 20 TOEFL test-related research projects are conducted by ETS's Research & Development staff each year and the TOEFL Committee of Examiners, comprised of language learning and testing experts from the global academic community, funds an annual program of TOEFL Family research by independent external researchers from all over the world.

The purpose of the *TOEFL Research Insight Series* is to provide a comprehensive yet user-friendly account of the essential concepts, procedures, and research results that help ensure the quality of scores for all members of the TOEFL Family of Assessments. Topics covered in these volumes feature issues of core interest to test users, including how tests were designed; evidence for the reliability, validity and fairness of test scores; and research-based recommendations for best practices.

The close collaboration with TOEFL score users, English language learning and teaching experts, and university scholars in the design of all TOEFL tests has been a cornerstone to their success and worldwide acceptance. Therefore, through this publication, we hope to foster an ever-stronger connection with our test users by sharing the rigorous measurement and research base and solid test development that continues to help ensure the quality of the TOEFL Family of Assessments.

Acknowledgements

The following ETS staff contributed to this version of Volume 4, updated in November 2024 (in alphabetical order): Larry Davis, Spiros Papageorgiou.

The following individuals also contributed to previous versions of this volume, by providing careful reviews and revisions, as well as editorial suggestions (in alphabetical order): Terry Axe, Ikkyu Choi, Michelle Hampton, John Norris, Eileen Tyson, and Yuan Wang. The primary authors of the first edition were Mary K. Enright and Eileen Tyson. Cristiane Breining, Rosalie Szabo, Xiaofei Tang, Mikyung Kim Wolf, and Xiaoming Xi also contributed to the first edition.

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 for proposed uses of tests" (American Educational Research Association, American Psychological Association®, & National Council on Measurement in Education [AERA, APA, & NCME], 2014, p. 11).

Test validation is the two-part process of first describing the proposed interpretations and uses of test scores and, second, investigating how well the test does what it is intended to do. Test validation thus starts by establishing an initial argument that states a series of propositions supporting the proposed interpretations and uses of test scores. It then involves posing questions for investigation, collecting data, and summarizing the evidence supporting these propositions (Kane, 2006, 2013). Because many types of evidence may be relevant, especially for high-stakes assessments, validation 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 it 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, read, 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 basic validity argument for the TOEFL iBT test, first by stating the propositions that underlie the proposed test score interpretations and uses and then by summarizing some of the evidence that has been found in relation to each proposition (see Table 1).

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

Proposition	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 field-study results for the test
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 tasks, grades, and other indicators of academic success
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)

Note. Another important proposition in the TOEFL validity argument, that test scores are reliable and comparable across test forms, is the subject of Volume 3 in this series.

In the following sections, we describe some of the main sources of evidence relevant to these propositions. The collection of this evidence for the TOEFL iBT test began with the initial discussions about a new TOEFL test in the early 1990s. These discussions prior to the design of the new TOEFL test led to many empirical investigations and evaluations of the results. 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. While a few highlights from this early validity research are summarized below, the bulk of the following focuses on more recent validity research that continues to monitor and update previous evidence, as well as to collect new evidence related to the uses of the TOEFL test.

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 to inform admissions decisions to English-medium colleges and universities, score users often want evidence that supports this proposition—evidence that the test content is authentic.

At the same time, it is important to emphasize that tests are events that are distinct from other academic activities. A single language test could never represent all of the types of language tasks that are encountered by students in the course of their university studies. Accordingly, test tasks and content—especially for large-scale standardized tests—are likely to be simulations and approximations, but never exact replications, of academic tasks. Accordingly, the TOEFL iBT test design process began with the analysis of real-life academic tasks and the identification of important characteristics of these tasks that could be captured in standardized test tasks that would function well with learners from around the world pursuing a wide variety of types of academic studies. This analysis focused on the general English knowledge, abilities, and skills needed to succeed in academic situations as well as the tasks and materials most typically encountered in colleges and universities. The development of the TOEFL iBT test also included 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, and Kunnan (2008).

Initial research that supported the development of relevant and representative test content included three empirical studies: Rosenfeld, Leung, and Oltman (2001); Biber et al. (2004); and Cumming, Grant, Mulcahy-Ernt, and Powers (2005).

Rosenfeld et al. (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 a broad range of reading, writing, speaking, and listening tasks were taken into consideration in the design of tasks for the TOEFL iBT test.

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. They also demonstrated constraints on the degree of authenticity that can characterize test tasks, due to the nature of what can and cannot be captured in a large-scale test setting. 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. It is a paramount concern that test content on the TOEFL iBT test be fair for all test takers. For this reason, unedited excerpts of authentic aural language from the corpus were not used as test materials. Many excerpts from the corpus required students to have knowledge other than that of the English language (e.g., mathematics), contained references to American culture that might not be understood internationally, or presented topics that might be upsetting to some students. Hence, the types of listening tasks represented in the corpus were used to model similar tasks in the TOEFL iBT test, while the authentic tasks themselves were not replicated in the assessment design.

One of the most innovative aspects of the TOEFL iBT test was the introduction of integrated test tasks—test tasks that require the integrated application of two or more language skills. Cumming et al. (2005) provided evidence about the content relevance, authenticity, and educational appropriateness of integrated tasks. Among the integrated test tasks included in the TOEFL iBT Speaking and Writing sections are some that require test takers to incorporate information from a brief lecture and a 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 their perceptions of the new tasks. The teachers viewed them positively, judging them 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' suggestions about how the tasks could be improved informed further refinement of the integrated task characteristics. In addition to integrated tasks, the TOEFL iBT test's Speaking and Writing sections also include independent test tasks that do not require the integration of information from Reading or Listening passages, instead asking test takers to express and explain personal preferences or choices.

Task Design and Scoring Rubrics

The design and presentation of tasks, and the rubrics (evaluation 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 multiple exploratory studies over 4 years to determine the best way to design new assessment tasks (Chapelle et al., 2008). These initial studies informed decisions about:

- Characteristics of the reading passages and listening materials
- Types of tasks used to assess reading and listening
- Types of integrated tasks used to assess speaking and writing
- Computer interface used to present the tasks
- Use of note-taking
- Timing of the tasks
- Number of tasks to include in each section

Careful attention was also paid to the development of rubrics (evaluation criteria) to score the responses to Speaking and Writing tasks. Groups of experts reviewed test takers' responses to pilot tasks and proposed scoring criteria. The rubrics were then trialed in field studies and revised, resulting in 4-point holistic rubrics for Speaking (ETS, 2023a), and 5-point holistic rubrics for Writing (ETS, 2023b). A similar process was used to establish the scoring rubric for the WAD task, which replaced the independent task in 2023. Unlike analytic rubrics in which various criteria for evaluation of a response are scored separately, holistic rubrics require the rater to consider all scoring criteria (e.g., delivery, language use, topic development) to produce a single holistic evaluation of the response. The original rubric development process was also informed by investigations of raters' cognitive processes as they analyzed test takers' responses (Brown, Iwashita, & McNamara, 2005; Cumming et al., 2006).

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, has been supported by multiple studies to date. These studies include investigations of the discourse characteristics of test takers' written and 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 scoring 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, and subsequently the Writing for Academic Discussion task in 2023, 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 of responses to these tasks varied as expected, both with writers' proficiency levels and with task types. The 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 and with 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. These findings have been replicated in recent studies that examined a larger number of responses (Knoch, Macqueen, & O'Hagan, 2014) and employed new measures of lexical sophistication (Kyle & Crossley, 2016). In addition, Plakans and Gebril (2017) analyzed 480 responses to integrated Writing tasks and found that, compared to responses that received low scores, high-scoring responses showed significantly better organization and cohesion. Davis and Norris (2023) analyzed the language produced by 242 test takers responding to both the independent writing task and the Writing for an Academic Discussion task. Measures of writing performances generated by automated natural language

processing tools revealed substantial similarities in the quality of texts produced by test takers on the two tasks, in terms of the syntactic complexity, grammatical accuracy, lexical variety, discourse cohesion and elaboration, and fluency of their writing. Findings provided initial support for replacing the independent writing task with the Writing for an Academic Discussion task to support interpretations about English writing ability.

For independent and integrated Speaking tasks, discourse analyses of responses to early prototypes were also carried out (Brown et al., 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 (forty 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. (2005) found that the qualities of spoken responses varied modestly with proficiency level and, to a lesser degree, 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. A study by Kyle, Crossley, and McNamara (2016) provides further evidence of the differences between test-taker responses to integrated and independent Speaking tasks. Using natural language processing tools, Kyle et al. showed that the independent tasks elicited less sophisticated words and more personal voice (pronouns and opinions) than the integrated tasks.

For reading tasks, an investigation of strategies used by test takers to answer 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 that are now used in the TOEFL iBT test. In summarizing the reading and test-taking strategies that were used for the full range of question types, the authors noted that test takers did not rely on test-wiseness strategies. Rather, according to the authors, 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 respond to 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, multicomponential 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 responses to 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 also consistent with the way test scores are reported and used (i.e., a total score and four skill scores). The 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). The invariance of the factor structure across different test-taker background variables has been further supported by recent factor analytic studies (Gu, 2014; Manna & Yoo, 2015; Sawaki & Sinharay, 2013), all pointing to desirable characteristics for how the test is structured.

Relationship Between TOEFL iBT Scores and Other Criteria of Language Proficiency

Another important proposition underlying valid score interpretation and use is that performance on the test is related to other indicators of or criteria for academic language proficiency. The central questions for test users are, "Does a test score really tell me about a student's performance ability beyond the test situation?" and "Is a student just a good test taker when it comes to the TOEFL iBT test? 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 to such questions 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 (GPA) in undergraduate or graduate studies often serves as a relevant criterion. However, the TOEFL test is intended to measure a more targeted 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 observed relationships: 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, or a restricted range, or an unusual distribution, 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, class attendance, and motivation. Thus, for example, correlations between similar types of measures are often quite high. Scores from the computer-based version of the TOEFL test, the iteration of the TOEFL test before the TOEFL iBT test (see TOEFL® Research Insight Series Volume 6: TOEFL® Program History), correlated very highly with scores from the TOEFL iBT test (observed r = .89, Wang, Eignor, & Enright, 2008). However, correlations between different types of measures, such as aptitude test scores and school grades, are typically more modest, on the order of r = .50 (Cohen, 1988).

With these caveats in mind, 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, academic placement, local institutional tests for international teaching assistants, performance on simulated academic listening tasks, and performance on real-world speaking and writing tasks.

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 et al., 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 more than 2,000 test takers completed the questionnaire. Observed correlations between the 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.

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 ESL development courses in addition to their content courses. Still other students may enroll themselves in intensive English programs (IEPs), hoping to improve their English language skills to prepare themselves for university study. These placements into ESL development courses and IEPs reflect a lower level of English language proficiency than unrestricted enrollment in content courses. In the 2003–2004 field study (Wang et al., 2008),

test takers who were studying in English-speaking countries were asked about their academic placement. Differences in test scores between students who were enrolled in ESL development courses or IEPs, and those enrolled in only content courses, were large and statistically significant, as illustrated in Figure 1.

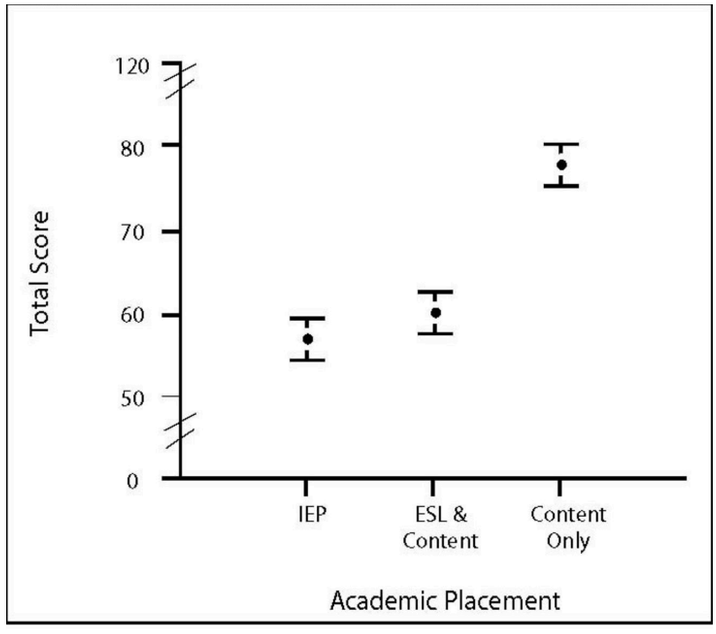


Figure 1. Mean total test score (+/- 95% CI) for test takers studying in English-speaking countries who were enrolled in institutional English programs (IEP, $n = 208$), taking both ESL courses and content courses ($n = 285$) or only content courses ($n = 488$).

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 prearrival screening measure for international teaching assistants (ITAs).
To this end, a standard-setting study (Wylie & Tannenbaum, 2006) established recommended cut scores for screening ITAs using TOEFL iBT Speaking section scores. Xi (2007, 2008) further investigated whether or not scores on the TOEFL iBT Speaking section could help institutions distinguish between candidates whose English was sufficient to begin teaching and those whose English was not. 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 distinct 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
Simulated teaching test (content and noncontent combined) scored on the basis of linguistic qualities (n = 84)	.78
Simulated teaching test (separate content- and noncontent-based tests) scored on the basis of linguistic qualities and teaching skills (n = 45)	.70
Simulated teaching test (content based) scored on the basis of linguistic qualities, teacher presence, and nonverbal communication (n = 53)	.53
Real classroom teaching sessions scored on the basis of linguistic qualities, lecturing skills, and awareness of American classroom culture (n = 23)	.44

Xi noted that the strength of the relationship was affected by the extent to which the local ITA tests engaged and evaluated nonlanguage abilities. The more the assessment focused on speaking abilities and the less on teaching skills, the higher 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. Wagner (2016) also found a relationship between TOEFL iBT Listening scores and students' perceptions of their ITAs' oral proficiency; this result was consistent with intended score interpretations, given that ITAs need to use their listening skills when they communicate with undergraduate students. Another volume in this series discusses the uses of TOEFL iBT scores to screen ITAs in greater detail (Educational Testing Service, 2021b).

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. To do so, Sawaki and Nissan first surveyed a sample of undergraduate and graduate students at four universities about the importance of a variety of academic listening tasks and course-related activities and assignments. This survey indicated that the most frequent and important 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 (2009) 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 disattenuated correlations between the TOEFL iBT Listening section scores and the simulated academic listening tasks were .62 for undergraduate students and .71 for graduate students. These findings suggest a substantial relationship between the TOEFL iBT Listening scores and the video-based academic lecture tasks.

Performance on Real-World Speaking and Writing Tasks

Similar studies were carried out with the criterion measures on academic speaking and writing abilities. Ockey, Koyama, Setoguchi, and Sun (2015) compared the TOEFL iBT Speaking scores of 222 English majors from a Japanese university to their performances on a local academic oral ability test, which included a three-member group oral discussion task, a picture and graph description task, and a prepared oral presentation task. The observed correlations between TOEFL iBT Speaking scores and the local task scores were .68 for oral presentation, .73 for picture and graph description, and .76 for the group oral discussion. The three local speaking tasks were also scored on seven different components. The strongest associations with TOEFL iBT scores were observed in the components of pronunciation (.63–.71), fluency (.59–.74), and vocabulary/ grammar (.50–.75). The component scores on interactional competence (.63), descriptive skill (.61), question and answer skill (.61), and delivery skill (.51) were moderately associated with TOEFL iBT Speaking scores.

Brooks and Swain (2014) compared the TOEFL iBT Speaking scores for thirty graduate students to their performances during real-life academic contexts, as reflected by recordings of an in-class and an out-of- class speaking activity (such as formal and informal presentations, paired or small group discussions). The grammatical, discourse, and vocabulary features in the participants' oral production on both TOEFL iBT test and real-life tasks were analyzed and compared. In the three contexts, there was some overlap in the use of connectives (e.g., furthermore, and, although), passivation (e.g., be satisfied), nominalization (e.g., simulation, propulsion), and vocabulary types, while in other measures (grammatical complexity, grammatical accuracy, use of speech organizers, use of questions, and the use of informal language) the three contexts were distinct.

Using similar linguistic analysis methodology, Riazi (2016) compared texts produced in TOEFL iBT Writing tasks and real-life academic writing assignments (including essays, reports, problem questions, explanations, and proposals) from 20 graduate students studying in five universities in Australia. Twenty linguistic and discourse features were analyzed, including syntactic (five variables), lexical (nine variables), and cohesion (six variables) features. Results of a series of repeated measures analysis of covariance (ANCOVA) procedures indicated that the texts produced in the TOEFL iBT Writing task and real-life academic assignments were similar on four out of the five syntactic features, five out of the nine lexical features, and all six cohesion features.

Malone and Llosa (2019) compared performance on two TOEFL iBT® writing tasks with performance in required writing courses for 103 international undergraduates studying in US universities. Performance on TOEFL iBT® writing tasks was evaluated in terms of task scores, while classroom writing performance was measured by writing instructors' ratings of student proficiency and grades on two course assignments. Additionally, both test and classroom writing were scored for five aspects of writing quality: grammatical, cohesive, rhetorical, sociopragmatic, and content control. Scores on the TOEFL iBT writing tasks were moderately and significantly correlated with teachers' judgements of students' writing ability and overall language proficiency. Also, scores on specific aspects of writing were significantly correlated across writing from test and classroom tasks.

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 higher educational institutions. For this particular decision and use, 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 opportunities to qualified students. Studies focused on U.S. universities—by Cho and Bridgeman (2012) and Bridgeman, Cho, and DiPietro (2016)—as well as a study of the British university context—Harsch, Ushioda, and Ladroue (2017)—show that there is a meaningful relationship between the TOEFL scores of students admitted to universities and their future academic performance indicated by their GPA. It is clear from this research that the higher a student's TOEFL score is, the higher the student's probability of demonstrating academic success in the form of GPA, especially during the first year of university study. This relationship was found to be meaningful because English language proficiency is a necessary (though not sufficient) condition for international students to succeed in universities where English is the medium of instruction. Other factors, such as subject-related knowledge and noncognitive attributes (e.g., motivation, persistence, and grit) can influence future academic performance as well.

Using test scores appropriately to make decisions with positive consequences is the joint responsibility of the test user and the test publisher. To support appropriate use of TOEFL iBT scores, ETS has provided score users with descriptive information to help them interpret test scores (ETS, 2021a; Wang & Papageorgiou, 2023), guidance on how to set standards for using scores at their institution for admissions purposes (ETS, 2021b), empirical studies, described above, on the effectiveness of speaking scores in making decisions about ITAs , and support decisions that are informed by external language proficiency levels and descriptors, in particular

particular those in the Common European Framework of Reference (CEFR; Council of Europe, 2001). Papageorgiou, Tannenbaum, Bridgeman, and Cho (2015) investigated the relationship between TOEFL iBT scores and the CEFR levels, and the TOEFL Steps research project further established this relationship (Educational Testing Service, 2023c). Research-based comparison tables between the TOEFL and IELTS® academic scores and between the TOEFL and CEFR are available at https://www.ets.org/toefl/institutions/ibt/compare-scores.

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 the 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. The manual Helping Your Students Communicate with Confidence (ETS, 2004) 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.

The impact of the TOEFL iBT test on teaching and learning has also been investigated in a multi-year research project. Wall and Horák (2006, 2008, 2011) studied how English language teachers in Eastern Europe coped with changes in the test, and whether and how test preparation materials changed in response to the new test. The investigation involved four phases:

- Phase 1 (Wall & Horák, 2006) constituted a baseline study in which observations were carried out and interviews were conducted with teachers, students, and directors at ten 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 & Horák, 2008) 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 & Horák, 2011) analyzed the coursebooks used by four of the teachers in Phase 2 as they continued to prepare students for the computer-based version of the TOEFL test and began to plan a course for the TOEFL iBT test. The analysis revealed that the TOEFL iBT test coursebooks differed considerably from the TOEFL CBT test coursebooks in terms of content, in reflection of the changes in test content and format. However, the coursebooks did not differ greatly in terms of their general methodological approach.

- Phase 4 (Wall & Horák, 2011) included observations of classroom teaching and interviews with a few of the Phase 1 teachers and directors of their institutions to observe what, if any, changes in teaching

occurred. While some aspects of teaching seemed not to have changed a great deal, others have changed considerably. Greater attention was paid to the development of speaking and there was a new focus on the integration of multiple skills.

Conclusion

Almost two decades of continuing research since the launch of the TOEFL iBT test has established a strong evidentiary case for the validity of proposed score interpretations and uses. Concerns about test validation were an integral consideration during the test design process, and the evidence gathered during that process has been comprehensively documented and synthesized (Chapelle et al., 2008). Beyond that foundational work, considerable additional evidence has been collected in response to important questions about the ways in which the TOEFL test is constructed, how examinees respond, and in particular, how scores are used and what consequences ensue. Certainly, test validation is an ongoing process that continues to be actively supported by ETS through the work of its own research staff and its funding of external research through the TOEFL Committee of Examiners (COE) research program (. The COE, composed of distinguished experts in language learning and assessment from the worldwide academic community, helps ETS orient TOEFL research in critical directions. It also publishes an annual announcement of a research program and invites language teaching and testing experts to submit proposals for TOEFL related research. In this way, the case for valid TOEFL score interpretation continues to grow and be refined.

References

Alderson, J. C. Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, 26(4), 621-631. doi:10.1177/0265532209346371

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *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., ...Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph No. 25). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, 33(3), 307–318.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly*, 11, 353–373.

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. 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, NY: Routledge.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL IBT® scores to academic performance: Some evidence from American universities. *Language Testing*, 29, 421–442.

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. 33). Princeton, NJ: Educational Testing Service.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

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 Monograph No. 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 the new TOEFL* (TOEFL Monograph No. 30). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2020a). *Guidelines for setting useful score requirements for the TOEFL iBT® test*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v9.pdf

Educational Testing Service. (2020b). *Reliability and comparability of TOEFL iBT scores.* https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v3.pdf

Educational Testing Service. (2021a). *Performance descriptors for the TOEFL iBT® test*. https://www.ets.org/pdfs/toefl/toefl-ibt-performance-descriptors.pdf

Educational Testing Service. (2021b). *Using TOEFL iBT® test scores for selecting international teaching assistants*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v10.pdf

Educational Testing Service. (2023a). *Scoring guides (rubrics) for TOEFL iBT Speaking responses*. https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

Educational Testing Service. (2023b). *Scoring guides (rubrics) for TOEFL iBT Writing responses*. https://www.ets.org/pdfs/toefl/toefl-ibt-writing-rubrics.pdf

Educational Testing Service. (2023c). *TOEFL® Steps: Building the learning path of the TOEFL family*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v11.pdf

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing,* 31(1), 111–133.

Harsch, C., Ushioda, E., & Ladroue, C. (2017). *Investigating the predictive validity of TOEFL iBT test scores and their use in informing policy in a United Kingdom university setting* (TOEFL iBT Research Report No. 30). Princeton, NJ: Educational Testing Service.

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* (pp. 55–96). New York, NY: Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). *An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT writing test* (TOEFL iBT Report No. 23). Princeton, NJ: Educational Testing Service.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, 33(3), 319–340.

Llosa, L., & Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing*, 36(2), 235-263.

Manna, V. F., & Yoo, H. (2015). *Investigating the relationship between test-taker background characteristics and test performance in a heterogeneous English-as-a-second-language (ESL) test population: A factor analytic approach* (Research Report No. RR-15-25). Princeton, NJ: Educational Testing Service.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, 32(1), 39–62.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). T*he association between TOEFL iBT test scores and the Common European Framework of Reference* (CEFR) levels (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing*, 31, 98–112.

Riazi, A. M. (2016). Comparing writing performance in TOEFL iBT and academic assignments: An exploration of textual features. *Assessing Writing*, 28, 15–27.

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. 08). Princeton, NJ: Educational Testing Service.

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (TOEFL iBT Research Report No. 21). 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. 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. 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* (pp. 27–54). New York, NY: Routledge.

Wagner, E. (2016). *A study of the use of the TOEFL iBT test speaking and listening scores for international teaching assistant screening* (TOEFL iBT Research Report No. 27). Princeton, NJ: Educational Testing Service.

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. 34). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2008). *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. 05). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3, The role of the coursebook. Phase 4, Describing change* (TOEFL iBT Research Report No. 17). Princeton, NJ: Educational Testing Service.

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* (pp. 259–318). New York, NY: Routledge.

Wang, L., & Papageorgiou, S. (2023). Scale anchoring methodology for developing revised performance level descriptors for the TOEFL iBT test. In S. Papageorgiou & V. F. Manna (Eds.) *Meaningful language test scores: Research to enhance score interpretation* (pp. 80-98). John Benjamins.

Wylie, E., & Tannenbaum, R. (2006). *TOEFL academic speaking test: Setting a cut score for international teaching assistants* (Research Memorandum No. RM-06-01). Princeton, NJ: Educational Testing Service.

Xi, X. (2007). Validating TOEFL iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly,* 4(4), 318–351.

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. 03). Princeton, NJ: Educational Testing Service.