ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–22
ETS Research Report No. RR–14-42

Stakeholders' Beliefs About the *TOEFL iBT*® Test as a Measure of Academic Language Ability

Margaret E. Malone

Megan Montee

December 2014

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

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

❖ ❖ ❖

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

Currently this research is carried out in consultation with the TOEFL Committee of Examiners (COE). Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from academia. The committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the TOEFL COE serve 4-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2014-2015) members of the TOEFL COE are:

Sara Weigle - Chair	Georgia State University
Yuko Goto Butler	University of Pennsylvania
Sheila Embleson	York University
Luke Harding	Lancaster University
Eunice Eunhee Jang	University of Toronto
Marianne Nikolov	University of Pécs
Lia Plakans	University of Iowa
James Purpura	Teachers College, Columbia University
John Read	The University of Auckland
Carsten Roever	The University of Melbourne
Diane Schmitt	Nottingham Trent University
Paula Winke	Michigan State University

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

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

ETS is an Equal Opportunity/Affirmative Action Employer.

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

TOEFL iBT Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Stakeholders' Beliefs About the *TOEFL iBT*® Test as a Measure of Academic Language Ability

Margaret E. Malone & Megan Montee

Center for Applied Linguistics, Washington, DC

The *TOEFL iBT*® test presents test takers with tasks meant to simulate the tasks required of students in English-medium universities. Research establishing the validity argument for the test provides evidence for score interpretation and the use of the test for university admissions and placement. Now that the test has been operational for several years, additional evidence is needed to support the validity argument, as well as to identify directions for future research or changes to the test. To address this need, this study examines the extent to which students, instructors, and university administrators understand and agree with the construct of academic language underlying TOEFL iBT tasks.

Keywords TOEFL iBT; validity argument; washback; focus groups; surveys; stimulated recall

doi:10.1002/ets2.12039

A validity argument is an organic and ongoing process, and the original evidence for the *TOEFL iBT*® test's validity, as discussed in Chapelle, Enright, and Jamieson (2008), provides solid groundwork for test score justification and use. Chapelle et al. emphasize that the test presents tasks that mirror the kind of academic tasks students will have to perform in the classroom. The test format differs from that of the TOEFL iBT's predecessors, the paper-based and computer-based forms, which tested listening, reading, and writing skills discretely. The TOEFL iBT presents test takers with integrated tasks that require the application of multiple skills to produce a response. For example, a test taker may have to read a passage, listen to a short lecture, and then produce an oral or written response. The Educational Testing Service (ETS, 2007) noted that "TOEFL iBT emphasizes integrated skills and provides better information to institutions about students' ability to communicate in an academic setting and their readiness for academic coursework" (p. 3).

The change in test item type represents a shifting definition of academic language ability. Whereas the TOEFL iBT 's predecessors defined academic language ability as the sum of a series of discrete skills (listening, reading, and writing), the TOEFL iBT presents academic language ability as the synthesis of discrete skills into an integrated whole. The TOEFL iBT thus purports to test the student's ability to carry out tasks that require such synthesis. In addition, the TOEFL iBT integrates speaking into the test's construct, unlike earlier versions of the *TOEFL*® test that assessed speaking through a separate test. Because the test tasks are intended to reflect real-world academic language, it is important to explore what stakeholders believe about the components of academic tasks and then to compare these to the actual structure of the TOEFL iBT tasks.

The test development process included two studies of stakeholder beliefs and perceptions of academic tasks (Cumming, Grant, Mulcahy-Ernt, & Powers, 2005; Rosenfeld, Leung, & Oltman, 2001). These studies informed the development of test tasks and integrated tasks in particular. The current study extends and expands this research by providing data from stakeholders who are affected by the operational use of the test. The purpose of this study is to gather information from key stakeholders about the TOEFL iBT as a measure of academic language ability. In order to provide a comprehensive picture of stakeholder beliefs about the test and its use, this study gathered information from three stakeholder groups: university administrators, instructors, and students. Multiple data collection methods, including focus groups and surveys of all three groups, and stimulated recalls with students were used to collect data about stakeholder attitudes about the test and test tasks and their relationship to real-world academic language use.

Three cultural contexts constituted the focus of the study. According to 2007 data, the three countries with the largest numbers of TOEFL iBT examinees in three major world areas, Europe, Asia, and the Middle East, were Germany, Korea,

Corresponding author: M. Malone, E-mail: mmalone@cal.org

and Saudi Arabia, respectively. The study included students from these three world areas who resided in either their home countries or the United States at the time of data collection. The focus groups included instructors in their home country and in the United States, while only U.S.-based instructors were included in the survey. Finally, all administrators included in the study worked at U.S. universities at the time of the study.

The results of the multiple-method study provide information about the extent to which users understand and agree with the construct that underlies TOEFL iBT tasks and the extent to which the beliefs of users abroad differ from those of users in the United States. Based on the results, recommendations are made for ways to improve test delivery and perceptions of the TOEFL iBT for stakeholders from specific world areas.

Background

Stakeholder beliefs about the correspondence between test tasks and the skills they purport to test provide important support for test validity. As mentioned previously, Chapelle et al. (2008) have articulated an extensive validity argument for the test, noting that inferences about test utilization are based on the assumption that the meaning of test scores is clearly interpretable by test users (p. 22). In addition to Chapelle et al.'s research, ETS (2008) states:

> 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. . . . [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. (p. 2)

Building from the test's validity argument and supported by ETS's public commitment to investigating user beliefs about the TOEFL iBT, this study examines three stakeholder groups: administrators, instructors, and students.

While research has been conducted on student and instructor beliefs about testing in general and the TOEFL and TOEFL iBT in particular (e.g., Cumming, Grant, Mulcahy-Ernt, & Powers, 2004; Hamp-Lyons, 2007; Hamp-Lyons & Brown, 2006; Shohamy, Donitsa-Schmidt, & Ferman, 1996; Stricker & Attali, 2010; Takagi, 2010; Wall & Horák, 2006), very little has been gathered about administrator beliefs. Because so little information was available about administrator beliefs, focusing on administrators as a stakeholder group presented both a challenge and an opportunity. The dearth of research left little to challenge or address, and at the same time, little on which to build. Therefore, conducting focus groups with this stakeholder group to ascertain as much information as possible prior to the large-scale survey was crucial to support development and administration of a survey that would elicit relevant data for the validity argument.

While little research exists on administrator beliefs, a larger body of research examines both instructor and test-taker (student) beliefs about tests. A number of studies (Cumming et al., 2004; Hamp-Lyons & Brown, 2006; Wall & Horák, 2006) have focused on instructor beliefs about the integrated tasks. In particular, Cumming et al. (2004) investigated instructor beliefs about the construct validity, authenticity, and educational relevance of the prototype integrated TOEFL tasks. Additional research has examined the impact of the TOEFL iBT on teaching and learning (Wall & Horák, 2008) through an in-depth study of English language teachers in Eastern Europe prior to and during the transition to the TOEFL iBT. Through this research, Wall and Horák (2008) conclude that teachers need to be prepared for changes in assessments, particularly in such a high-stakes and far-reaching test as the TOEFL iBT. Hamp-Lyons (2007) supported these outcomes by pointing out that any new assessment efforts are unlikely to yield the expected results without preparing instructors for these changes; the TOEFL iBT is an excellent example of such an innovation. While Kern (1995) discerned that teacher beliefs are but one factor that can effect student results, it is nonetheless true that teacher beliefs do have an impact on student beliefs. Therefore, investigating teacher beliefs, and recognizing these beliefs may have an impact on student results, is a cornerstone of this study.

Shohamy et al. (1996) emphasized that a discrepancy often exists between test takers' and test creators' beliefs about what a test actually measures. Research on student beliefs about the TOEFL iBT has focused on a variety of issues, including student attitudes toward computer-based testing and beliefs about what is being measured. The research on student attitudes toward computer-based testing exhibited by students taking the computer-based TOEFL test (Jamieson, Taylor, Kirsch, & Eignor, 1999; Stricker, Wilder, & Rock, 2004) indicates positive beliefs on the part of students toward the computer-based TOEFL (although not the computer-delivered iBT). Stricker and Attali (2010), in an extensive study of

test-taker attitudes toward the TOEFL iBT, found that attitudes differed by world area and test section. They found that students from Germany were either neutral or negative toward the TOEFL iBT, while students from China, Columbia, and Egypt held more positive views toward the test. Moreover, Stricker and Attali found that test-taker attitudes were most positive toward the listening and writing sections and least positive toward the speaking sections. While previous TOEFL research provides a useful perspective on stakeholders' attitudes toward the test, changes to the recent form of the test necessitate further investigation into this aspect of test validity. This study will examine its outcomes in relationship to Stricker and Attali's results.

As previous research and as ETS's promotional materials attest, one way to investigate the validity issue is to consider the beliefs of key stakeholder groups. For this study, as previously stated, stakeholders include administrators, instructors, and students. University administrators responsible for admissions and placement decisions hold beliefs about how effectively the TOEFL iBT represents students' English skills and whether or not students are effective communicators in English-speaking (specifically U.S.) universities. Similarly, instructors hold beliefs about both how effectively the TOEFL iBT tests English skills and whether students are effective communicators in English-speaking (specifically U.S.) universities and instruction in English is consistent with preparation for the test. Finally, students hold beliefs about whether the TOEFL iBT measures their English ability and whether the abilities tested by the TOEFL iBT are consistent with what is needed for an English-speaking university. These beliefs may or may not depend on the countries from which students originate and whether they are currently studying in a university whose medium of instruction is English. However, research into these beliefs is needed to determine to what extent administrator, instructor, and/or student beliefs support the validity argument.

Beliefs, Perceptions, and Attitudes

The importance of research on stakeholder beliefs, perceptions, and attitudes about tests is widely recognized. However, the definitions of these terms overlap with one another and, at times, contradict one another in the literature. In this study, researchers use the term *belief* and will adopt Hamp-Lyons and Brown's (2006) definition of the term: "Beliefs refer to the stakeholders' perceptions about the various dimensions of test preparation practice . . . and frequently have an overt or unconscious ideological element" (p. 19). Beliefs are characterized as underlying conscious or unconscious perceptions that entail some sort of value judgment on the part of the individual expressing that belief.

Research on Instructor Beliefs

Research on the testing of second language acquisition has repeatedly demonstrated that test tasks should reflect real-life learning situations (Bachman, 1990, 2002; Darling-Hammond, 1994; Freedman, 1991; Linn, Baker, & Dunbar, 1991). A number of studies have suggested that teacher input can improve testing (Brindley, 1998, 2001; Chalhoub-Deville, 1995; Elder, 1993; Epp & Stawychny, 2001; Grant, 1997; North, 1995, 2000; O'Sullivan, Weir, & Saville, 2002; Stansfield & Kenyon, 1996; Teachers of English to Speakers of Other Languages [TESOL], 1998). In addition, Griffin (1995), Hoge and Coladarci (1989), Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett (2001), and Sharpley and Edgar (1986) have suggested that teacher input can provide benefits for test design, because such input often results in suggestions for a more authentic learning situation reflected in the test.

One study (Cumming et al., 2004) has looked at instructor attitudes toward the integrated tasks on the TOEFL iBT as a part of a larger study on score reliability and student performance on the TOEFL iBT (Enright & Cline, 2002). Cumming and colleagues interviewed seven English as a second language (ESL) instructors at three universities to determine whether the content of the tasks reflected the domain of academic English utilized in the North American university setting, whether students' performance on prototype test items reflected their language use in the classroom, and whether the tasks were aligned with the claims that ETS had made about their design. According to the researchers, most respondents believed that the tasks reflected the academic English used in their classes and that 70% of student responses on prototype items were consistent with students' language use in the classroom. However, the data for this study were limited to interviews with a relatively small number of instructors; data from a larger number and greater variety of TOEFL stakeholders would help provide a clearer picture of beliefs about this test.

Research on Student Beliefs

To date, little research has examined student test-taker perceptions of the items on the TOEFL iBT. Stricker et al. (2004) looked at the attitudes toward computer-based testing manifested by students taking the computer-based TOEFL, and other research on students has focused exclusively on test takers' attitudes toward computer-based testing (Powers & O'Neill, 1993; Schmitt, Gilliland, Landis, & Devine, 1993; Schmidt, Urry, & Gugel, 1978). Rosenfeld et al. (2001) conducted extensive studies of listening, speaking, reading, and writing tasks necessary for academic success; while their research is illuminating, it does not directly reflect the content of the TOEFL iBT.

Research on Washback

Washback is commonly defined as the changes that occur in teaching and learning as a result of test use (Hamp-Lyons & Brown, 2006). However, Hamp-Lyons and Shohamy (2004) have suggested that impact, rather than washback, is a more accurate way of referring to the influence a test may have not only on education but also on society. Certainly, the iBT — administered to 246,120 examinees between September 2005 and December 2006 — has the potential to effect change within both educational and greater societal contexts.

Alderson and Hamp-Lyons (1996), in a study on the washback effects of TOEFL preparation courses, investigated claims that the TOEFL has a negative influence on language teaching and concluded that definitions of washback and hypotheses of how it works need to be more complex. They found that teachers' attitudes toward teaching TOEFL preparation courses were generally negative. In addition, their analysis of classroom observations found substantial differences between behaviors observed in the TOEFL classes and those observed in the non-TOEFL ESL classes. However, Hamp-Lyons and Brown (2006) found that all respondents (students and teachers) believed that the TOEFL was a good predictor of ability to cope in a university setting.

Wall and Horák (2006) conducted a three-stage study to record a baseline description of teaching and learning in Central and Eastern European TOEFL preparation courses prior to the introduction of the iBT. The purpose of this study was to determine the impact of the introduction of the test on the same users across three studies. In the first stage, they surveyed the test creators to obtain a better understanding of the *positive impact* that the new test was intended to have on classroom design. They found that TOEFL iBT creators agreed that the revisions in the test were intended to have a positive effect on the classroom by encouraging an increased emphasis on communicative approaches to teaching and by promoting work on integrated, rather than discrete, skills.

In short, the literature suggests that stakeholder beliefs can contribute to or detract from a test's validity argument. Therefore, the focus of this project was to investigate the beliefs of three groups of stakeholders: administrators, instructors, and students (test takers). Beliefs were investigated via multiple methods. First, the researchers conducted focus groups with students and instructors in the United States and overseas, as well as with U.S. university administrators from admissions and international student offices to identify common themes among each group in their beliefs about the TOEFL iBT. In turn, the themes from the focus groups were used to inform the development of the surveys. Finally, stimulated recalls were conducted with students. Figure 1 shows the different contexts for stakeholder groups.

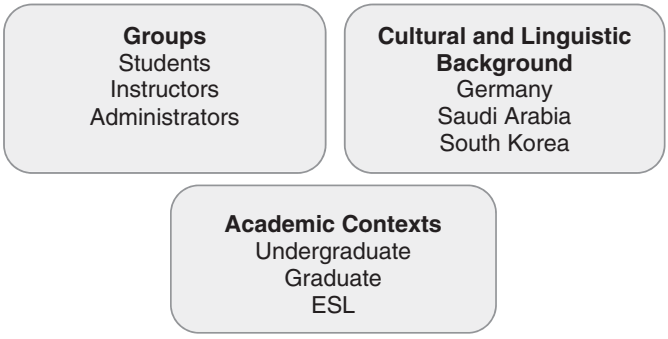

Figure 1 Contexts for stakeholder groups.

As Figure 1 indicates, the study sought to elicit beliefs from a variety of contexts. Germany, Saudi Arabia, and South Korea were chosen because each has the highest total of TOEFL iBT examinees tested between September 2005 and December 2006 in its region (Europe, the Middle East, and Asia, respectively). Participants in Germany, Saudi Arabia, and South Korea included only native speakers of German, Arabic, and Korean, respectively.

The student group contained the largest number of participants and included students from all three cultural and linguistic backgrounds (home country and United States) in the focus groups, surveys, and stimulated recalls. Because of difficulty recruiting Saudi students in the United States, the focus groups and stimulated recalls included students from other Arabic-speaking countries. In the instructor group, participants from both the United States and target countries were included in the focus groups, but only instructors from the United States participated in the survey. Finally, administrators from the United States participated in the focus groups and surveys.

The study investigated five research questions via multiple methods: focus groups, surveys, and stimulated recalls.

1. What are users' beliefs about what the TOEFL iBT measures? This research question was addressed in the focus groups and surveys of administrators, instructors, and students.
2. How do users define academic language ability in their contexts? This research question was addressed in the focus groups and surveys of administrators, instructors, and students.
3. Do student beliefs differ based on context? This research question was explored via student surveys.
4. How well do users believe integrated TOEFL iBT tasks measure students' academic language ability? Do those beliefs vary by context? This research question was explored via the surveys, as well as through stimulated recalls conducted with students.
5. Do users believe that student performances on integrated TOEFL iBT tasks reflect performance in academic classes? Do these beliefs vary by context? This research question was explored via instructor and student surveys.

Focus Groups Study

This study relied on focus groups to generate topics for the subsequent large-scale survey of users. Focus groups provide one way to generate information from the target audience and allow a setting in which participants can exchange ideas on the focus group topics until the session reaches the point of information saturation (that is, when no new insights are emerging). Focus groups elicit opinions and beliefs and can generate data that would not come up in individual interviews or very structured settings (Mackey & Gass, 2005). In addition, focus groups are an efficient means of gathering data from multiple interviewees in less time than individual interviews.

For this study, focus group methodology was particularly useful because it allowed for an open exchange of ideas among a small group of stakeholders. Focus group results informed subsequent phases of the study but were not used to directly answer the research questions. For this initial part of the study, focus groups were used as an exploratory tool to investigate the following questions:

1. What are users' beliefs about what the TOEFL iBT measures?
2. How do users define academic language ability in their contexts?

This focus group study used a multiple-category design in order to gather data from more than one group of stakeholders (Krueger & Casey, 2000). This approach allowed for comparisons within and across categories. In other words, although separate focus groups were held for each group of stakeholders (administrators, instructors, or students) within a specific world area (United States, Germany, Saudi Arabia, or South Korea), the topics developed so that the results could be cross-referenced and compared within and across groups. This meant that the researchers could determine similarities and differences within student and teacher groups, whether held in the United States or overseas, as well as similarities and differences across stakeholder groups of administrators, instructors, and students. This approach also meant that the survey questions could be designed to maximize future comparisons of responses. That is, if overlap occurred among groups in terms of topic, the surveys could be designed to develop complementary questions for future comparison.

While a focus group format can provide rich data, the researchers had to ensure that all participants had sufficient experience with the TOEFL iBT to provide such information in a focus group setting. All focus group participants

Table 1 Participants and Number of Focus Groups by Country and User Group

	Germany	Saudi Arabia	South Korea	United States	United States
Administrators	N/A	N/A	N/A	3 (1)[a]	4 (1)[b]
Instructors	6 (1)	5 (1)	6 (1)	5 (1)	
Students	6 (1)	7 (1)	6 (1)	6 (2)[c]	4 (1)[d]

Note. The study involved only one group of instructors from the United States.
[a]Admissions officers. [b]International student affairs officers. [c]Students enrolled in an ESL program at a U.S. university. [d]Students enrolled in an academic program at a U.S. university.

completed a screening questionnaire to demonstrate familiarity with the TOEFL iBT in general and the integrated tasks in particular (Appendix A). Screening criteria included test familiarity, TOEFL iBT score (for students), teaching experience (for instructors), and professional responsibilities (for administrators). The research team developed a focus group protocol in collaboration with overseas consultants in Germany, Saudi Arabia, and South Korea. Consultants then translated the final protocols into German, Arabic, and Korean. Focus groups were conducted in the United States and in the overseas contexts. All focus groups in the United States were conducted in English, and the student groups included participants from all three cultural contexts. Focus groups in Germany, Saudi Arabia, and South Korea were conducted in the students' native languages. Table 1 provides a breakdown of participants by region and user group. The total number of participants for each category is listed with the number of focus groups in parentheses. For example, there were two focus groups conducted with ESL students in the United States and six total participants in these two groups.

As Table 1 shows, students, instructors, and administrators participated in the focus groups. Student and instructor groups included participants in Germany, Saudi Arabia, South Korea, and the United States; all student participants in the United States were from the cultural groups targeted in the study. However, because of difficulty recruiting students from Saudi Arabia in the United States, this focus group was expanded to include two students from other Arabic-speaking countries.

The purpose of the focus groups was to gather information about participants' attitudes toward the TOEFL iBT and identify topics to explore in subsequent phases of the study, specifically to develop questions for the large-scale stakeholder survey. Appendix A includes selected questions from the student and instructor focus groups. For these groups, a protocol was used for all focus groups to ensure comparability across contexts. For the administrator focus group, it was determined that a more open-ended, topical format would be appropriate because the research team had less information about this user group and their beliefs about the TOEFL iBT. Because more literature was available about student and instructor beliefs, the questions for these focus groups were more targeted than those for the administrators.

Procedures

Researchers followed a system of open coding evolving from the data to reflect themes that emerged in the groups. During each focus group, a moderator and observer took notes individually, and the sessions were audio recorded. After the focus groups were conducted, the moderator and observer reviewed their notes individually and listened to the audio recordings to revise and add to their notes as necessary.

The next step of coding involved a synthesis of themes. Instructors and students fell into two major groups: those who participated in the United States and those from overseas. Administrators also fell into two groups: those who worked in admissions and those who worked with international students. Therefore, the moderator and observer met to review the topics that emerged for each group — students, instructors, and administrators — to establish the most common topics and to ensure that all topics were represented. After a first step of identifying the major topics for each group, the moderator and observer referred to their notes from the focus group sessions to select portions of the audio files for transcription. For the overseas groups, these portions were then translated into English by the consultant.

All focus group notes and transcripts were coded by members of the research team at the Center for Applied Linguistics (CAL). The researchers coded the materials for thematic categories using an emergent coding scheme. Table 2 lists the major themes that emerged from each group (administrators, instructors, and students) as a result of this iterative process.

Table 2 Major Focus Group Themes by Group

	Administrators	Instructors	Students, ESL	Students, university
Preparation for the test	Relation to proficiency When should students prepare	Test format vs. language learning	Washback from pBT to TOEFL iBT	Authenticity Relevancy for United States
Administration of the test	Delivery Test center technology	Delivery Test center technology	Delivery Test center technology	Delivery Test center technology
The iBT	Familiarity with the TOEFL iBT Familiarity with changes	Integrated tasks Length Time pressure Familiarity with the TOEFL iBT Familiarity with changes	Topics Integrated tasks Time-pressure	Topics Content knowledge vs. English ability Academic content vs. social situations Integrated tasks Time-pressure
Speaking sections on the iBT	Authenticity Time constraints	Authenticity Time constraints	Authenticity Time constraints	Authenticity Time constraints
Cultural issues	Preparation methods and test scores Bias for some populations	Preparing for and taking the test	Bias for some populations	Cultural differences in test preparation
Impact of the test	Change in student population because of test requirements Tracking students to show impact/success of students	Communication with ETS Level of sophistication Tracking students to show impact/success of students	Score as a reflection of English ability	Score as a reflection of English ability

Note. iBT = Internet-based test; pBT = paper-based test; ESL = English as a second language, ETS = Educational Testing Service.

Table 3 Illustrative Quotes From Administrators and Instructors

	Administrators	Instructors
Preparation for the test	Test preparation is needed. In order to be on the same level playing field, they have to know about the test.	The delicate balance that we have as instructors to teach both language and test wiseness and not letting the test wiseness overtake the language, but if we only focus on language, they won't be prepared for the TOEFL.
Administration of the test	A lot of students can't type very fast — so their scores are lower than would be expected — bias.	Change the order — start with the reading and then the writing, then break, then do the speaking and the listening.
The iBT	[The] new edition is wonderful. Tasks make sense because [they are] closer to the actual thing that students have to do.	Students who come in at a higher level do the best … the ones with the highest language ability to start got the highest score. Others with lower ability are having difficulty.
Speaking sections on the test	[Students] need to know about the TOEFL; it's a weird test. [In the speaking section] you've got 15 seconds to prepare something and speak for 45 [seconds].	Incredible frustration level for the speaking. [Students have] difficulty collecting their thoughts that quickly.
Cultural issues	Cultural issues on the test — how students prepare and how well they do — is there bias in the test?	[There is a] difference between countries that students are coming from — Asians felt good with grammar. Latin Americans felt good with speaking.
Impact of the test	I don't feel that it has infiltrated the international community yet […] Only the best students had the desire to take the iBT.	Integrated tasks foment the need for synthesis. [It's the] ultimate skill to be able to do college level research.

As Table 2 indicates, similarities and differences are present across the groups and contexts. Among all groups and contexts, test administration, the speaking portion of the test, and test impact emerged as central themes. Similar issues emerged across all groups regarding the administration of the test in regards to concerns about test delivery format and test center technology. In discussing speaking, all groups highlighted the lack of authenticity and time constraints. While there were differences in topics among groups on the impact of the test, both administrators and instructors expressed concern about using test scores to track students to show their successes. Similarly, students in both the ESL and university classes were concerned about the TOEFL iBT score as a reflection of English language ability. Consistencies were also found between instructors and administrators in the topics generated by the focus groups. Both groups discussed their level of familiarity with the TOEFL iBT and raised concerns about not being aware of specific details about the test.

At the same time, the differences in topics raised by the groups were also interesting. One noteworthy outcome was that the administrator focus groups demonstrated a lack of basic familiarity with the TOEFL iBT and with the changes made from the paper-based TOEFL. While instructors also expressed some unfamiliarity with the TOEFL iBT, their lack of familiarity with the changes emerged as a lack of familiarity with the content of integrated tasks, rather than with the changes in general; in addition, instructors who taught TOEFL preparation classes, as would be expected, demonstrated the most familiarity of all instructors with the specific changes from the paper-based TOEFL and the TOEFL iBT. Another interesting outcome stemmed from the topic of cultural issues. Both ESL students and administrators identified potential bias toward specific populations as an issue with the TOEFL iBT, while the instructors and university students focused on test preparation as a possible cultural difference. In general, the topics raised by the focus groups provided important information for the content of the survey. First, the results of the focus groups suggested that the administrators would need more background to explain changes to the TOEFL iBT and would likely have less insight into these changes. Second, the results of the focus groups identified areas to be highlighted in the surveys for each group and across groups. Finally, the focus groups provided rich insight into the beliefs of each group. Tables 3 and 4 present sample quotes from each group on each topic to illustrate the issues that emerged in the focus groups.

Table 4 Illustrative Quotes From ESL and University Students

	ESL students	University students
Preparation for the test	Because the test has four sections, it was more difficult to prepare for it than other tests. But it was nice because I had to study four skills of English that were needed linguistic factors to study in the United States.	The class that best prepared me for the test is the writing section—write essays and then give to professor who would return corrected. Never in your life you would have to do that again—artificial—and need to prepare.
Administration of the test	But when I think about the reading section, I remember most that next to me and behind me other people were doing their speaking sections, and it confused me so much and it was so difficult to concentrate. That part is really poorly managed. You can barely concentrate.	[It's a] bad environment—everyone goes through and on sections at different time.
The iBT	iBT is more about skills than structure.	It would be better [if] it focused on academic things rather than our experience—topics are too social.
Speaking sections on the test	I thought the time limit for the speaking part was actually the worst. I just watched the timer count down and then thought, hmm, and now what do I say? I need to hear a voice. I can't speak for 2 minutes to someone who's sleeping.	Speaking part—why can't you talk to someone rather than recording it? It would make it easier for students.
Cultural issues	[Saudi students] can't talk to a computer. We want to talk to a person.	Largely, the attitudes are a reflection of the culture. Spelling is a huge problem because of the lack of vowels. TOEFL is great for groups like Asians who love test taking. Arabs are used to taking more like a European battery test that is oral.
Impact of the test	This test helped me to improve my writing ability, but it does not seem to measure students' writing ability that they can use in the academic context in the United States. It is short.	TOEFL limited because [the test] doesn't have higher academic vocabulary.

Note. ESL = English as a second language; iBT = Internet-based test.

As Tables 3 and 4 demonstrate through representative quotes, many opinions were similar across groups and contexts. In developing the surveys and stimulated recalls, CAL researchers included focus group themes in the survey for the appropriate context (administrators, instructors, and students). Each of the six major themes that emerged from the focus groups were integrated into multiple questions on the survey. In addition to informing survey questions about stakeholder attitudes, the focus group data also provided important information for questions about background information, including specific information about stakeholder familiarity with the TOEFL iBT and specific changes to the test.

The next section describes the research questions for the major part of the study, the survey of administrators, instructors, and students.

Survey

The development of the survey questions was informed by a review of research and the results of the stakeholder focus groups. Because this study focused on stakeholder beliefs about the TOEFL iBT, it examined the extent to which users believe that tasks mirror actual academic tasks that students in the U.S. university system encounter. Participants included university administrators, instructors, and students (test takers). Data were collected from university administrators in the United States; ESL instructors in Germany, Korea, Saudi Arabia, and the United States; university students who are

native speakers of German, Korean, or Arabic (Saudi dialect) in their home countries; and university students who are native speakers of German, Korean, or Arabic in the United States. Survey data were used to address all five of the study's research questions and form the core of the results. Data from the stimulated recalls are discussed separately.

Survey Instrument Development

The next phase of the study involved developing the three online surveys, one each for administrators, instructors, and students, to collect large-scale data on user beliefs about TOEFL iBT items. Drawn from the major themes gathered from the focus groups, a literature review about the TOEFL, and a review of the TOEFL iBT preparation materials and test items, each survey targeted a specific user group: students, instructors, or university administrators.

The administrator surveys included 24 questions and focused on testing and admissions criteria. Given the limited body of research about the beliefs of university administrators about the TOEFL, a survey of administrators conducted as part of research about the International English Language Testing System (IELTS; Hawkey, 2006) informed the development of the administrator survey. The instructor surveys included 40 questions and focused on the structure of the classes they taught, as well as the extent of the instructors' interaction with the TOEFL iBT. The student survey was translated into the three target languages (see Appendix D for an English version). Each student survey asked 42 questions and included sample TOEFL iBT integrated tasks for survey participants' reference to ensure that all survey participants had a clear example of an integrated task.

Surveys were administered using Survey Monkey, an online delivery format. Appendices B, C, and D include the survey questions for each user group (the student surveys include the English version). Before being released, the surveys were reviewed by native and nonnative speakers of English, as well as the TOEFL Committee of Examiners. After revision, the student surveys were translated into Arabic, German, and Korean. For the Arabic survey, Modern Standard Arabic was used. Each survey was piloted with 20 respondents and revised a final time, including the translated version, before being released.

Instructor and University Administrator Participants

The instructor and university administrator surveys were written in English and sent to respondents in the United States. University administrators were contacted in several ways, including announcements in professional LISTSERVs and direct e-mail messages to university administrators. Instructors were reached through announcements in professional LISTSERVs and professional organizations, direct e-mail messages to private and university-based English language programs, and social networking websites including Twitter and Facebook. Because of the indirect methods used to reach participants (e.g., Facebook), it is not possible to calculate the response rate for instructors and administrators. Table 5 shows the number of surveys collected from instructors and administrators.

Table 5 shows both the number of surveys collected, or the number of participants who responded to at least part of the survey, and the number completed by each group. In order to be included in the completed survey group, administrator and instructor participants had to complete the background section and 90% of all survey questions. When data are missing through nonresponse of the participants, it is noted in the result tables.

Sampling Frame for Student Participants

For the student surveys, CAL sent an e-mail invitation to participate, translated into the appropriate target language, to a student sample provided by ETS. The sample included students who had completed the TOEFL iBT within an 18-month period. Proportional numbers of graduate and undergraduate students were represented, spread evenly across low-, mid-,

Table 5 Number of Surveys Completed and Collected From Administrators and Instructors

Context	*N* completed (collected)
Administrators	246 (295)
Instructors	289 (434)
	Total 535 (729)

Table 6 Number of Surveys Completed and Collected From Students

Context	*N* completed (collected)
Germany	279 (313)
Saudi Arabia	274 (331)
South Korea	235 (266)
	Total 788 (910)

and high-range scores. The survey sampling deviated from the planned distribution because the students' location at the time of sampling could not be determined (either abroad or in the United States). To reach the target of 250 surveys per cultural group, survey invitations were sent via e-mail to 1,500 students per group (German, Korean, and Saudi). Of the students contacted by e-mail messages provided by ETS, the total response rate was 17%, which was lower than the expected response rate of 20%. The low response rate may be due to difficulties in sending e-mail translated into the target languages or in e-mailing students with non-English-language accounts. Across all student surveys, 85% of respondents who began the survey completed all questions. Incomplete surveys were not analyzed. As a participation incentive, students were entered in a raffle to win a $200 gift certificate.

In order to reach the target number of students, social networking Web sites including Twitter and Facebook were also used to distribute the survey. However, these networks yielded minimal responses ($N = 78$); these responses were not used in calculating the response rate. Table 6 lists the number of surveys collected from students in each group.

The data in Table 6 show that the target sample of student participants was exceeded for surveys collected and completed ($N = 788$). For students from Germany and Saudi Arabia, the targeted N, 250, was exceeded; for South Korean, the number of completed surveys was only 15 participants shy of the targeted number.

The next section addresses the extent to which surveys from each group represent the expected population. In addition to collecting information about stakeholder beliefs about the TOEFL iBT, the surveys for each group—administrators, instructors, and students—also collected information about the respondents' backgrounds. These questions were designed to find out the kinds of institutions with which respondents were affiliated and their experiences working with the TOEFL.

Survey Analysis

The five research questions were primarily addressed by descriptive statistics. For Research Questions 3, 4, and 5, inferential statistics were used to compare survey responses between groups. To control the overall error rate of the study, a Bonferroni correction was applied with an overall alpha level of .05, which established the alpha level for each of the 11 inferential tests at $\alpha = .005$. Effect size is reported as Cohen's d for t-tests with $d = .20$, .50, and .80 interpreted as small, medium, and large effect sizes, respectively (Cohen, 1988). For analyses of variance, effect size is reported as η^2 with $\eta^2 = .01$, .06, and .14 as small, medium, and large effect sizes, respectively (Cohen, 1988).

Administrator Background

The surveys elicited a great deal of information about the administrators' background. This section reports on the most relevant pieces of information, including institutional information and number of students served per year.

Institutional Information

All respondents to the administrator survey were asked to report their institutional information in terms of (a) types of institutions, (b) state, and (c) number of international students their institutions served each year. The respondents reported coming from 47 states, each of which is represented by 1 – 17 respondents. Table 7 shows the kinds of institutions with which the respondents were associated.

As Table 7 indicates, the majority of the respondents work in private or public, 4-year colleges or universities. In addition, most participants served more than 200 students per year.

Table 7 Types of Institutions and Number of International Students Served Each Academic Year

Institutional information	Response categories	Frequency	%
In which type of institution do you currently work?	Trade or vocational school	0	0
	Community or junior college	8	3
	Private college or university (4-year)	120	49
	Public college or university (4-year)	100	41
	Other	18	7
Approximately how many international students does your institution serve each academic year?	100 international students or fewer	94	38
	101 – 200 international students	38	15
	201 – 500 international students	39	16
	501 – 1,000 international students	32	13
	1,001 international students or more	36	15
	Don't know	7	3

Note: $N = 246$, no missing responses.

The survey also asked administrators about the types of activities they perform relative to international students. The most common responses (86% and 81%, respectively) were "responding to admissions questions" and "giving input for international admissions policies"; 72% said that they reviewed applications, while 62% recruit and 65% advise admitted students. More than half (61%) make decisions about international admissions policies. Few administrators reported much interaction with international students for placement (24%) and even fewer for helping build students' English language skills (16%). Therefore, the data show that most administrators who responded to the survey were primarily involved in tasks related to admission, recruitment, and advising of international students, as well as working with university admissions policy.

English Language Proficiency Requirements

Administrators were asked about their institutions' English language proficiency requirements in terms of whether or not their language requirements had changed in the past 2 years, which English tests were endorsed, and what alternatives were accepted in place of an English test score. Table 8 shows the results from this question.

As Table 8 shows, 98% of the respondents reported accepting the TOEFL iBT as a test of English language proficiency; 95% reported accepting TOEFL pBT, and 76% reported accepting IELTS. Other tests that their institutions accepted included, for example, Michigan English Language Assessment Battery (MELAB) and *SAT®*; 42% reported accepting alternatives to test scores. In summary, administrators from 47 of the 50 U.S. states participated in the survey, and nearly all (98%) report accepting the TOEFL iBT for admission. The next section examines the background of instructors who responded to the survey.

Table 8 Tests Accepted to Meet English Language Proficiency Requirements

Which of the following tests does your institution accept to meet entrance requirements for English language proficiency?	Frequency	%[a]
TOEFL iBT	241	98
TOEFL pBT	236	96
IELTS	186	76
MELICET	37	15
TSE	27	11
TOEIC®	24	10
STEP	22	9
London Test of English	7	3
Other	37	15

Note. $N = 246$, no missing responses. IELTS = International English Language Testing System; MELICET = Michigan English Language Institute College Entrance Test; TSE = Test of Spoken English; STEP = Standardized Test for English Proficiency.

[a]Respondents were able to select more than one answer; therefore, totals exceed 100%.

Table 9 Instructors' Positions and Instructional Settings

	Response Categories	Frequency	%[a]
Which best describes your current position?	Instructor	174	60
	Private tutor	9	3
	Both an instructor and a private tutor	55	19
	Other	51	18
Where do you currently teach?	University-based ESL/EFL institution	171	59
	Private ESL/EFL institution	81	28
	Self-employed	26	9
	Tutoring company	14	5
	Non-ETS test preparation center	12	4
	ETS test preparation center	10	4
	Other	39	14

Note. $N = 246$, no missing responses. ESL = English as a second language; EFL = English as a foreign language; ETS = Educational Testing Service.

[a]Respondents were able to select more than one answer; therefore, totals exceed 100%.

Instructor Background

Instructor Characteristics

Characteristics of the instructor sample were compiled in terms of instructor background, including ESL/English as a foreign language (EFL) teaching experience, and teacher preparation; ESL/EFL pre-service/in-service training; position/current work setting; and program type.

English as a Second Language/English as a Foreign Language Teaching Experience and Instructor Preparation

All respondents to the instructor survey were asked about their ESL/EFL teaching experience and types of training they received in teaching ESL/EFL. The results to the survey showed that instructors had a great deal of experience teaching the target population. Of the total instructors, 92% reported teaching ESL/EFL for 3 years or more, and 60% of the instructors had taught ESL/EFL for 7 or more years. Of the total respondents, 62% received a preservice university degree related to teaching, and 49% reported receiving an ESL/EFL teaching certificate. Of the total respondents, 20% reported receiving training other than the kinds listed in the survey.

Positions and Instructional Settings

Of the total respondents, 60% reported working as instructors, and 19% reported working as both an instructor and a private tutor. Table 9 provides detailed results.

Table 9 shows the instructor positions and instructional settings; 18% reported holding positions other than the listed categories. These included positions such as academic coordinator, administrator, and curriculum developer, and joint positions such as administrator and instructor or instructor and academic advisor. Of the respondents, 59% reported working in university-based ESL/EFL institutions, and 28% reported working in private ESL/EFL institutions. Reponses to the *other* category indicated that 14% of the respondents worked in community colleges or government agencies; 9% of the respondents reported self-employment. The majority of the respondents (76%) described their programs as a program for students who wish to enter a university. Of the respondents, 29% described their programs as a program for students who are currently enrolled in non-ESL university classes (Table 8); 22% described their programs as *other*. The next section describes the student population that responded to the survey.

Student Background

Table 6 shows the number of students who submitted and completed the survey. Overall, 910 students from Germany, South Korea, and Saudi Arabia participated in the survey; 788 students completed the survey for an 87% completion rate

Table 10 Highest Level of Education and Current Level of Study

		Frequency			%		
	Response categories	G	SA	SK	G	SA	SK
Highest level of education	High school/secondary school diploma	125	71	89	46	27	38
	College/undergraduate degree (AB, BA, BS)	67	144	99	25	54	42
	Graduate degree (master's, doctoral, etc.)	61	41	42	22	15	18
	Other	19	12	4	7	5	2
	Total *N* (missing responses)	272 (7)	268 (6)	234 (1)			
Current level of study	High school/secondary school	30	2	19	11	1	8
	College/undergraduate level	56	89	107	21	33	46
	Graduate level (master's, doctoral, etc.)	109	135	58	40	50	25
	I am not a student.	35	39	42	13	15	18
	Other	72	8	8	27	3	3
	Total *N* (missing responses)	302 (0)	273 (1)	234 (1)			

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.

overall. The number of surveys completed for students from Germany ($N = 279$) and Saudi Arabia ($N = 274$) were similar, while students from South Korea ($N = 235$) were the fewest.

The survey sample included test takers from a variety of proficiency levels. Survey respondents were asked to self-report their TOEFL iBT scores and subscores using ranges. Overall, survey respondents reflected general trends in TOEFL iBT scores by country, as reported by ETS (2010), with students from Germany tending to report higher scores than students from Saudi Arabia and South Korea. The mean score for German test takers in 2009 was 96 out of a total possible score of 120. The mean score for South Korean test takers was 81 and was 62 for test takers in Saudi Arabia.

On the survey, 74% of German respondents reported a score of 90 or above, and 30% reported a score of 110 or above. Students from South Korea tended to report lower TOEFL iBT scores than German respondents. Korean respondents reported a variety of scores on the TOEFL iBT but had fewer students at the highest end of the scale; 19% of Korean respondents reported scoring 110 or higher on the test. Finally, Saudi respondents tended to report lower TOEFL iBT scores than the other groups, with very few students (6%) reporting scores above 100. A total of 63% of Saudi respondents reported TOEFL iBT scores of 79 or below. Thus, the survey generally reflected overall trends in TOEFL iBT scores by country.

Characteristics of all three student groups were compiled in terms of level of education and prior learning of English; prior experience with the TOEFL iBT; and TOEFL iBT preparation class. The next few sections show and compare the results by group.

Respondents were asked about their highest level of education, current level of study, number of years learning English, prior exposure to English, ESL or academic courses that they were and/or are currently enrolled in, and the country where they are studying or studied English and where English is the primary language. No questions were asked about age. Table 10 shows the highest level of education and current level of study for students from Germany, Saudi Arabia, and South Korea.

As Table 10 indicates, respondents from the three contexts had different backgrounds. More than half (54%) of the Saudi students who responded had completed an undergraduate degree, and 42% of students from South Korea had an undergraduate degree. By contrast, one quarter of German students who responded to the survey (25%) had completed an undergraduate degree. Some German students also indicated that they were enrolled in both undergraduate and graduate level courses simultaneously; therefore, the totals for this question added up to more than 100%. Students were also asked about their current level of study; half of the students from Saudi Arabia were graduate students, compared with 40% of German students and 25% of South Korean students. Overall, more graduate students ($N = 302$) than undergraduates ($N = 252$) responded to the survey. Student respondents also answered questions about their previous English language learning. Table 11 shows the amount of time and grade levels when students studied English.

Table 11 shows that almost all German students (90%) and 68% of South Korean students who responded had studied English for 6 or more years. This percentage contrasted sharply with students from Saudi Arabia; 43% of students from Saudi Arabia who responded had studied English for only 1 year, whereas only 28% had studied English for more than 6 years. The most interesting and possibly conflicting information came from students from Saudi Arabia,

Table 11 Prior Learning

		Frequency			%[a]		
	Response categories	G	SA	SK	G	SA	SK
How long have you studied English?	1 year		116	7		43	3
	2 years	2	46	16	1	17	7
	3 years	4	3	14	2	1	6
	4 years	1	9	12	0	3	5
	5 years	5	8	15	2	3	6
	6 years	15	11	11	6	4	5
	More than 6 years	245	75	159	90	28	68
	Total *N* (missing responses)	272 (7)	268 (6)	234 (1)			
Did you study English in:	Kindergarten (age 3–6)	6	77	30	2	29	13
	Primary school (age 7–11)	75	94	109	28	35	47
	Secondary school (age 12–17)	267	255	225	98	95	96
	College/university	192	228	199	71	85	85
	Other	33	28	8	12	10	3

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.

[a]Students could respond *yes* to as many as applied; therefore, the percentages are more than 100.

who reported results that were inconsistent with the number of years they reported studying the language. Although 43% reported only having studied English for 1 year, 29% reported having studied English in kindergarten, 35% in primary school, and 95% in secondary school. This suggests that students from Saudi Arabia may not have understood the question.

The survey also asked students about the amount of exposure to English that respondents had in their English classes. The question specifically asked students about the amount of time they spent reading English texts in their chosen fields. Responses showed that students reported different experiences in hearing English in class from their teachers and reading English texts. German students reported being spoken to by their teachers least in English during class and reading the fewest English texts in their chosen field, and Saudi students had similar responses. More than 50% of students from South Korea reported that their instructors spoke English often or always. This suggests that classes are conducted differently in the different countries. On the other hand, students from all contexts reported seldom reading English language texts in their field of study. This suggests that, while the amount of English spoken by instructors is different in South Korea than in Germany or Saudi Arabia — or that students' perceptions about frequency of spoken English differs from country to country — few students from any contexts rely on English language texts.

The survey next asked students about their experience in studying English in an English-speaking country. Overall, nearly half (48%) of respondents were currently studying in an English-speaking country, but more Saudi students (68%) were studying in an English-speaking country at the time of the survey than South Koreans (46%) and Germans (31%). The 48% of students previously or currently studying in an English-speaking country were also asked about what they had studied or were studying in that country. German students who had previously studied in an English-speaking country reported most frequently having taken academic language classes over the students from Saudi Arabia and South Korean students. In contrast, more South Koreans reported being currently enrolled in academic classes over the students from Saudi Arabia and Germany.

Students were also asked to indicate their motivation for taking the TOEFL iBT; Table 12 shows how students reported their reasons for taking the test.

As Table 12 shows, the primary reason students reported taking the TOEFL iBT was that it was the only form available. The next most common reason was that the TOEFL iBT was required by the program. Saudi students reported a strong preference for a computer-based test compared with German and South Korean students.

Students also responded to a question about the other English tests they had taken. Table 13 shows the results.

Table 13 indicates that country of origin was important in determining which other tests students had taken. For example, nearly 80% of German students reported that they had taken no other or similar test, South Korean students reported the TOEIC as the next most common test taken, and Saudi students reported the IELTS as the next most popular test among their survey participants.

Table 12 Reasons for Taking TOEFL iBT

	Frequency			%[a]		
Why did you take the Internet form (iBT) of the TOEFL?	G	SA	SK	G	SA	SK
It was the only form available.	144	153	134	53	57	57
I prefer computer tests to other available testing modes.	54	83	13	20	31	6
My program or university only accepts iBT scores.	74	82	83	27	31	36
I received my scores faster by taking the Internet-based test.	45	51	13	17	19	6
The iBT shows my language abilities better than other TOEFL forms.	15	58	18	6	22	8
I was prepared in school or class to take the Internet-based version only.	13	43	23	5	16	10
Other	24	12	6	4	5	3

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.
[a]Respondents were able to select more than one answer; therefore, totals exceed 100%.

Table 13 Tests Similar to TOEFL iBT

		Frequency			%[a]		
	Response categories	G	SA	SK	G	SA	SK
Have you taken tests similar to the TOEFL iBT?	IELTS	4	75	10	2	28	4
	London Test of English	0	3	0	0	1	0
	MELICET	2	30	2	1	11	1
	STEP	3	0	3	1	0	1
	TOEFL pBT	6	34	38	2	13	16
	TOEIC	2	6	99	1	2	42
	TSE	2	3	6	1	1	3
	I did not take a similar test.	215	134	100	79	50	43
	Other	39	12	16	14	5	7

Note. G = Germany; SA = Saudi Arabia; SK = South Korea; IELTS = International English Language Testing System; MELICET = Michigan English Language Institute College Entrance Test; STEP = Standardized Test for English Proficiency; TSE = Test of Spoken English.
[a]Respondents were able to select more than one answer; therefore, totals exceed 100%.

The survey also asked about student preparation for the TOEFL. The results to this question show the differences that country of origin had on TOEFL preparation. German respondents were least likely to take a TOEFL preparation course (16%), followed by students from Saudi Arabia (46%), and then South Korean students (65%).

Results

This section addresses the heart of the study: What do users believe the TOEFL iBT measures? The results are described by research question and group (administrators, instructors, and students).

Research Question 1: What Do Users Believe the TOEFL iBT Measures?

Administrators and instructors responded to a number of statements regarding their understanding of and background regarding the TOEFL iBT.

Administrator Beliefs About the TOEFL iBT

All respondents to the administrator survey were asked to respond to statements about their understanding of and beliefs about the TOEFL iBT. Tables 14 and 15 summarize the administrators' beliefs about what the TOEFL iBT measures. For this part of the survey, administrators rated the extent of their agreement with statements about the TOEFL iBT according to a 4-point scale:

- 4 (*strongly agree*)

Table 14 Administrator Background on the TOEFL iBT

Please indicate your level of agreement with the following statements:	*N*	*Mean* rating[a]	*SD*	No response
I am confident using TOEFL iBT scores to make admissions decisions.	212	3.09	.58	34
Candidates have a better chance of getting a good score on the TOEFL iBT if they attend a preparation course.	158	3.08	.59	66
I am familiar with how international students prepare to study in an American university.	221	2.98	.72	25
I understand how to interpret TOEFL iBT scores.	210	2.93	.73	36
The TOEFL iBT is a good predictor of how well international students will perform at my institution.	200	2.90	.61	46
I understand how the TOEFL Internet-based test is different from the TOEFL paper-based test (pBT).	211	2.87	.82	35
The test publisher disseminates adequate information about the meaning of TOEFL iBT scores.	179	2.83	.70	67
I am familiar with the content of the TOEFL iBT.	224	2.79	.79	22
I look at both composite and subscores on the TOEFL iBT.	196	2.76	.91	50
Candidates need to prepare for the TOEFL iBT using materials designed specifically for the TOEFL iBT.	146	2.76	.70	100
I am familiar with the way international students prepare for the TOEFL iBT.	214	2.62	.73	32
Students who have taken the TOEFL iBT have better language skills than those who have not.	160	2.46	.73	86

[a]Minimum = 1; maximum = 4.

Table 15 Administrator Beliefs About the TOEFL iBT

Please indicate your level of agreement with the following statements:	*N*	*Mean* rating[a]	*SD*	No response
Students from some cultural groups have higher scores on the TOEFL iBT than students from other cultural groups.	178	3.21	.64	68
The TOEFL iBT is a good measure of English language proficiency at the undergraduate level.	214	3.14	.46	32
The TOEFL iBT is a good measure of English language proficiency at the graduate level.	177	3.08	.52	69
The TOEFL iBT is a good measure of English language proficiency at the pre-university level.	158	3.06	.60	88
The content of the TOEFL iBT reflects what students need to be able to do at a university in an English-speaking country.	179	3.01	.55	67
The TOEFL iBT is a good measure of English language proficiency for vocational studies.	105	2.91	.64	141
International students' TOEFL iBT scores are similar to my own perceptions of their language ability.	196	2.85	.54	50
The TOEFL iBT is a fair measure of English language ability for all populations of international students.	180	2.84	.65	66
International students' TOEFL iBT scores relate to their academic achievements in university courses.	169	2.78	.57	77

[a]Minimum = 1; maximum = 4.

- 3 (*agree*)
- 2 (*disagree*)
- 1 (*strongly disagree*)

As the literature indicates, there is no definitive standard on how to determine the midpoint for data of this kind (McIver & Carmines, 1981). Clearly, a mean of 3.0 or higher indicates agreement, and a mean of 2.0 or lower indicates disagreement. Therefore, the researchers examined the data and determined that a mean agreement of 2.75 or higher indicates that most respondents agreed with the statement; below that number means that, on average, respondents did

not agree with the statement. Table 14 shows the extent to which administrators agree with specific statements about the TOEFL iBT. Table 14 and subsequent tables list statements in order of mean level of agreement.

As Table 14 shows, mean ratings by administrators about their background on the TOEFL iBT indicate that they agree with most of the statements above. However, the average rating for the statements "I am familiar with the content of the TOEFL iBT" and "I look at both composite and subscores on the TOEFL iBT" are only barely above the 2.75 level, as is the statement regarding whether candidates need to prepare for the TOEFL iBT using specifically designed materials. In addition, the results show that these administrators do not indicate familiarity with how students prepare for the TOEFL iBT. Finally, administrators do not necessarily consider students who have taken the TOEFL iBT to have better language skills than those who have not; however, it is important to note that the differences among all the ratings is fairly small.

Table 15 shows what administrators believe about specific statements regarding the TOEFL iBT.

As Table 15 shows, administrators agree with all of the provided statements and believe that the TOEFL iBT is a good measure of English ability. At the same time, many statements have a number of nonresponses, and average responses to many statements approach the 2.75 cutoff for *agree*. For example, on average, administrators rated the statement "International students' TOEFL iBT scores are similar to my own perceptions of their language ability" 2.85 and, similarly, the statement "International students' TOEFL iBT scores relate to their academic achievements in university courses" at 2.78. In addition to the borderline agreement, a number of administrators did not respond to these statements. Combined, this suggests a relative lack of enthusiastic agreement with the statement as compared to more highly, on average, rated statements such as "The content of the TOEFL iBT is a good measure of English language proficiency at the undergraduate level." It is also possible that these ratings and the missing responses reflect administrators' lack of familiarity with the content of the TOEFL iBT.

Instructor Beliefs About the TOEFL iBT

All respondents to the instructor survey were asked about their beliefs about the four sections in the TOEFL iBT. For this part of the survey, instructors rated the extent of their agreement with statements about the TOEFL iBT according to the 4-point scale previously described; a mean of 2.75 was established as an indicator of agreement. Table 16 shows instructors' beliefs about the TOEFL.

As Table 16 indicates, instructors somewhat disagreed with all statements about the TOEFL, from understanding what the subscores mean to ETS's effectiveness in disseminating information about the TOEFL and its recent changes. No statement has a mean rating at or above 2.75, indicating that instructors tended to disagree with the statements. Notably, instructors tended to disagree that preparing to take the TOEFL iBT prepares students for life at an English-speaking university or that the test predicts how well students will perform at an English-speaking university. In addition, instructors reported a low agreement (under 2.0) that they understand what TOEFL iBT scores mean. Instructors also believe that users only look at the total and not the subscores. Perhaps most importantly, instructors suggested that ETS could do more to disseminate information about the TOEFL iBT. Finally, instructors did not agree that they examine the subscores in making decisions about student scores.

Table 16 Instructor Beliefs About the TOEFL iBT

Please indicate your level of agreement with the following statements:	*N*	*Mean* rating[a]	*SD*	No response
The TOEFL iBT is an accurate predictor of how well a nonnative English speaker will perform at an English-speaking university.	226	2.65	.72	63
Users of TOEFL iBT scores understand how to use TOEFL iBT scores.	175	2.63	.79	114
Preparing to take the TOEFL iBT prepares students for life at an English-speaking university.	242	2.31	.77	47
ETS adequately disseminates information about the meaning of TOEFL iBT scores.	201	2.11	.68	88
ETS adequately disseminates information about changes to the TOEFL iBT.	194	2.06	.68	95
Users of TOEFL iBT scores look at subscores as well as total scores.	198	2.01	.73	91
I understand what the TOEFL iBT scores mean.	263	1.93	.73	26

Note. ETS = Educational Testing Service.

[a]Minimum = 1; Maximum = 4.

Table 17 Instructor Beliefs About What the TOEFL iBT Measures

Please indicate your level of agreement with the following statements:	*N*	*Mean* rating[a]	*SD*	No response
The writing section allows students to show how well they can write in English.	259	3.20	.54	30
The listening section allows students to show how well they can listen in English.	261	3.11	.59	28
The speaking section allows students to show how well they can speak in English.	254	3.01	.62	35
The reading section allows students to show how well they can read in English.	259	3.07	.49	30

[a]Minimum = 1; Maximum = 4.

Table 18 Relevance by User Population

Is the TOEFL iBT appropriate to students' future English language needs:	*N*	Min	Max	*Mean*	*SD*
At the preuniversity level	206	0	1	.62	.49
At the undergraduate level	233	0	1	.85	.36
At the graduate level	215	0	1	.78	.42
For vocational studies	157	0	1	.46	.50

Table 17 shows instructor beliefs about what the TOEFL iBT measures.

Despite low average ratings about instructor beliefs about the TOEFL iBT overall, mean ratings on all four questions regarding the TOEFL iBT as a measure of domain-specific language ability were above 3.00, indicating that, on average, the instructor group agreed that the listening, speaking, reading, and writing sections allow students to show how well they can use English. Differences between the instructor and administrator beliefs will be reported with responses to Research Questions 4 and 5. Note that instructors basically agreed with all statements, although the speaking and reading sections had slightly lower levels of agreement than writing and listening.

Instructors also exemplified their beliefs by responding to yes/no statements made about the TOEFL iBT. Table 18 shows instructors' beliefs about the TOEFL iBT's effectiveness in a number of possible settings; because the responses were yes/no, a mean response of .75 indicates agreement. This is because 1.0 would indicate complete agreement, while 0 would indicate complete disagreement. An average of .5 would indicate a mean between agreement and disagreement. Therefore, .75 was selected as a conservative estimate of mean agreement. When asked whether the TOEFL iBT appropriately served students' future English language needs, instructors felt it was most appropriate at the undergraduate and graduate levels. Overall, instructors agreed that the TOEFL was appropriate at the undergraduate (*Mean* = .85) and the graduate levels (*Mean* = .78) and less appropriate at the preuniversity level and for vocational studies (*Mean* = .62 and .46, respectively).

Student Beliefs About the TOEFL iBT

For this part of the survey, students rated the extent of their agreement with statements about the TOEFL iBT according to the 4-point scale previously described; 2.75 was set as the standard for agreement with the statement. Table 19 shows what students from the three different countries believe about what the TOEFL iBT measures.

Table 19 shows the extent of student agreement with statements about the TOEFL iBT. These results indicate that students had mixed beliefs about whether or not the TOEFL iBT allowed them to show how well they can write, speak, read, and listen in English. As a group, German students agreed with all statements about the TOEFL iBT's ability to show how well they could perform in English, except on the speaking section. The German students were the only participants who agreed that the test questions felt natural. Students from all countries believed that the listening section showed how well they could listen in English, as the means were above 2.75 in all cases. Clearly, students did not believe that the TOEFL iBT allowed them to show their ability to speak English. Overall, Saudi and South Korean student responses indicated some disagreement with the TOEFL iBT's capacity to show their abilities in English.

Summary: Research Question 1

This section has demonstrated that instructors agree with most statements about the TOEFL iBT's ability to show how well students can read, write, listen, and speak in English, while most students do not agree that all sections of the TOEFL

Table 19 Efficacy of TOEFL iBT

Please indicate your level of agreement with the following statements:		*N*	*Mean* rating[a]	*SD*	No response
The test questions on the TOEFL iBT felt natural.	Germany	279	3.04	.59	0
	Saudi Arabia	274	2.42	.81	0
	South Korea	235	2.62	.60	0
The writing section on the TOEFL iBT let me show how well I can write in English.	Germany	279	2.82	.70	0
	Saudi Arabia	274	2.67	.79	0
	South Korea	235	2.74	.62	0
The listening section on the TOEFL iBT let me show how well I can listen in English.	Germany	279	3.19	.62	0
	Saudi Arabia	274	2.91	.78	0
	South Korea	235	2.92	.60	0
The speaking section on the TOEFL iBT let me show how well I can speak in English.	Germany	279	2.49	.81	0
	Saudi Arabia	274	2.62	.87	0
	South Korea	235	2.43	.80	0
The reading section on the TOEFL iBT let me show how well I can read in English.	Germany	279	2.91	.78	0
	Saudi Arabia	274	2.27	.91	0
	South Korea	235	2.84	.62	0

[a]Minimum = 1; Maximum = 4.

Table 20 Skills Needed for Language Domains

Reading	Listening	Writing	Speaking
• Finding the main idea • Organizing information • Summarizing a passage • Finding the relationship between ideas	• Taking notes • Finding the main idea • Finding the speaker's purpose • Drawing conclusions based on what is implied • Making connections between pieces of information in a conversation or lecture	• Using vocabulary appropriately • Organizing a cohesive essay • Following spelling conventions • Using a range of grammatical expressions • Using idiomatic expressions appropriately • Identifying one main idea and some supporting points	• Express opinions on topics • Summarize information verbally • Respond to questions in a timely manner • Talk about thoughts in an organized way • Participate in speech similar to an academic discussion

iBT measure their ability to use English. Finally, administrators agree that the TOEFL iBT does measure student ability and agree somewhat that it is a good predictor of student success in effectively using English in academic courses.

Research Question 2: How Do Users Define Academic Language Ability in Their Contexts?

The next research question addressed how the three groups (administrators, instructors, and students) defined academic language ability within their own contexts. This definition was established by first examining the TOEFL Framework and data from the focus groups. Then, the researchers developed survey questions to elicit administrator, instructor, and student beliefs about what comprised academic language ability However, the term *academic language ability* is a technical term used for specific purposes within educational contexts. Therefore, the survey did not directly ask respondents, "How do you define academic language ability?" Instead, it was necessary to allow users to define academic language ability indirectly, by having respondents answer questions on the frequency with which they used specific skills when responding to tasks on the TOEFL iBT or when preparing for the TOEFL iBT. These skills, listed in Table 20, emerged from both a review of the TOEFL iBT framework (Jamieson, Eignor, Grabe, & Kunnan, 2008) and from focus group data.

The researchers consulted the list of skill components from the TOEFL Framework (Chapelle et al., 2008) to inform the development of descriptors that comprise reading, listening, writing, and speaking. In reviewing this list and comparing it to the results of the focus groups, the researchers developed an initial survey that included nearly all of the skills from the framework. However, survey piloting indicated that the questions needed to be streamlined to reduce respondent fatigue. Therefore, the researchers consulted the focus group data to ensure that the skills listed within each domain on

Table 21 Task Use for the Reading Section of the TOEFL iBT

Approximately how often do you incorporate the following tasks into your instruction?	*N*	*Mean* rating[a]	*SD*	No response
Scan text for facts	288	3.60	.67	1
Read to understand vocabulary	287	3.57	.69	2
Infer how ideas connect	288	3.55	.69	1
Summarize reading passages	288	3.55	.69	1
Find the relationships between ideas	286	3.55	.66	3
Organize information	289	3.46	.72	0
Practice grammar structures	287	2.99	.88	2

[a]Minimum = 1; Maximum = 4.

the survey reflected what instructors and students thought was important and to ensure that survey questions would be easily understood by respondents. As a result, survey questions reflected the skills specified by the TOEFL Framework (Chapelle et al., 2008) and were further winnowed by focus group results to both ensure fidelity to the construct and comprehensibility by stakeholders who responded to the survey.

While instructors were relatively neutral on whether the TOEFL iBT, as a whole test, was an effective measure of student ability, they did agree that individual sections of the TOEFL iBT allowed students to demonstrate their skills in that domain.

Therefore, the answer to the second research question was determined by examining instructor and student responses to statements about how instructors prepare students and how students prepare themselves for the TOEFL iBT.

Instructor Responses

Instructors used a 4-point Likert scale to rate the degree of their agreement with statements about how they prepare students for the TOEFL by focusing on the skills listed in Table 20. Table 21 represents what instructors believe to be important in preparing students for the reading section of the exam.

The table above shows that instructors use all skills suggested for reading, with grammar being used least often and scanning the text being used most often in preparation. The relatively low rating of grammar structures does not suggest that a focus on grammar is not central in preparing students for the TOEFL iBT as the mean is still higher than 2.75. Table 21 suggests that all skill use is important in preparing students for the TOEFL iBT reading section.

Table 22 shows how instructors prepare their students for the listening section of the TOEFL.

Table 22 shows that instructors reported using all the tasks put forth on the survey to prepare students. The means are all close, ranging from 3.54 to 3.60, suggesting that instructors believe all these skills are important to prepare for the TOEFL iBT listening section.

Table 23 shows how often instructors report using specific tasks to prepare their students for the speaking portion of TOEFL iBT.

In Table 23, instructors reported using most of the tasks listed in preparing students for the TOEFL iBT. Two tasks that they rarely reported using were *finding opportunities for students to speak with native speakers* and *practicing formulas for responding to TOEFL questions*, as shown in Table 23. The data suggest that instructors do not find opportunities for students to conduct conversations with native speakers in order to prepare for the TOEFL. Instructors also did not report that formulas were important in preparing students for the TOEFL iBT, suggesting that these language instructors

Table 22 Task Use for the Listening Section of the TOEFL iBT

Approximately how often do you incorporate the following tasks into your instruction?	*N*	*Mean* rating[a]	*SD*	No response
Draw conclusions based on what is implied in the listening	289	3.60	.64	0
Listen for speaker's purpose	289	3.57	.68	0
Understand the relationships between ideas	289	3.57	.64	0
Listen for introductions, topic changes, and conclusions	289	3.57	.68	0
Take notes	289	3.56	.71	0
Make connections among pieces of information in a conversation or lecture	289	3.54	.67	0

[a]Minimum = 1; Maximum = 4.

Table 23 Task Use for the Speaking Section of the TOEFL iBT

Approximately how often do you incorporate the following tasks into your instruction?	*N*	*Mean* rating[a]	*SD*	No response
Read about a topic and then talk about it	289	3.48	.77	0
Make a point and provide supporting examples	289	3.44	.72	0
Give oral presentation	289	3.40	.82	0
Listen to a lecture or conversation and then talk about it	289	3.30	.83	0
Hold group discussions/debates	289	3.25	.88	0
Practice timed speaking	289	3.22	.88	0
Pronunciation	289	3.17	.79	0
Find opportunities for students to speak with native speakers of English	289	2.69	1.03	0
Practice formulas for responding to questions on the TOEFL iBT	289	2.69	1.08	0

[a]Minimum = 1; Maximum = 4.

Table 24 Task Use for the Writing Section of the TOEFL iBT

Approximately how often do you incorporate the following tasks into your instruction?	*N*	*Mean* rating[a]	*SD*	No response
Organize a cohesive essay	289	3.71	.61	0
Build vocabulary	289	3.57	.65	0
Write opinion essays	289	3.53	.72	0
Practice formula for structuring an essay	289	3.52	.75	0
Use a range of grammatical features	289	3.46	.73	0
Write about what students have learned from a listening or reading passage	289	3.36	.81	0
Follow spelling conventions	289	3.19	.81	0
Use idiomatic expressions	289	3.08	.72	0

[a]Minimum = 1; Maximum = 4.

believe that what is expected on the speaking section is not formulaic language but spontaneous speech. Therefore, it is possible that instructors use global tasks to prepare students for the test, with the exception of opportunities to speak with native speakers. Similarly, instructors reported timing their students' speaking but not practicing formulas for speaking. Based on this information, instructors' actions in preparing their students for the TOEFL iBT speaking section speak to somewhat mixed beliefs about what is important.

The final question about preparation for individual sections of the TOEFL dealt with writing. Table 24 examines what instructors reported doing most frequently to prepare students for the writing section.

As Table 24 shows, instructors reported using all tasks in preparing students for writing. While slight differences exist, these differences are all within the range of agreement.

Student Responses

This section discusses student responses to the survey questions about what skills were important to taking the TOEFL iBT. Table 25 shows how students from the three language groups responded to the skills they used to respond to the reading section of the TOEFL.

Table 25 Skill Use on the Reading Section of the TOEFL iBT

	N (no response)				*Mean* rating[a]				*SD*			
On the reading section of the TOEFL iBT, how often did you use the following skills?	G	SA	SK	Total	G	SA	SK	All	G	SA	SK	All
Find the main idea	267 (12)	263 (7)	234 (1)	764	3.41	3.46	3.45	3.44	.74	.65	.65	0.68
Organize information	263 (19)	266 (4)	233 (2)	762	3.25	2.88	3.11	3.08	.78	.86	.74	0.79
Find the relationships between ideas	261 (9)	259 (11)	233 (2)	753	2.95	2.78	3.05	2.93	.86	.86	.84	0.85
Summarize a passage	266 (13)	265 (5)	233 (2)	764	2.72	2.58	2.93	2.74	.93	.96	.85	0.91

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.

[a]Minimum = 1; maximum = 4.

Table 26 Skill Use on the Listening Section of the TOEFL iBT

On the listening section of the TOEFL iBT, how often did you use the following skills?	*N* (no response)				*Mean* rating[a]				*SD*			
	G	SA	SK	Total	G	SA	SK	All	G	SA	SK	All
Find the main idea	273 (6)	271 (3)	235 (0)	779	3.34	3.45	3.40	3.40	.72	.66	.72	.70
Find the speaker's purpose	268 (11)	270 (4)	235 (0)	773	3.14	3.54	3.39	3.36	.82	.65	.76	.74
Take notes	277 (2)	272 (2)	234 (1)	783	3.23	3.26	3.29	3.26	.95	.90	.86	.90
Make connections between pieces of information in a conversation or lecture	266 (11)	264 (10)	232 (3)	762	3.23	3.23	3.11	3.19	.67	.78	.74	.73
Draw conclusions based on what is implied	263 (16)	264 (10)	234 (1)	761	3.04	3.24	3.12	3.13	.79	.80	.76	.78

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.
[a]Minimum = 1; maximum = 4.

Table 27 Skill Use on the Writing Section of the TOEFL iBT

On the writing section of the TOEFL iBT, how often did you use the following skills?	*N* (no response)				*Mean* rating[a]				*SD*			
	G	SA	SK	Total	G	SA	SK	All	G	SA	SK	All
Organize a cohesive essay	276 (3)	267 (7)	232 (3)	775	3.47	3.30	3.26	3.34	.77	.75	.73	0.75
Follow spelling conventions	272 (7)	256 (18)	231 (4)	759	3.51	3.12	3.30	3.31	.72	.77	.69	.73
Use vocabulary appropriately	272 (7)	268 (6)	232 (3)	772	3.54	3.16	3.19	3.30	.65	.70	.62	.66
Identify one main idea and some supporting points	270 (9)	269 (5)	232 (3)	771	3.37	3.31	3.20	3.29	.72	.77	.79	.76
Use a range of grammatical expressions	272 (7)	262 (12)	231 (4)	765	3.15	3.13	2.90	3.06	.81	.76	.77	.78
Use idiomatic expressions appropriately	270 (9)	263 (11)	232 (3)	765	2.84	3.11	2.71	2.89	.94	.81	.83	.86

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.
[a]Minimum = 1; maximum = 4.

All of the skills represented in Table 25 were rated highly, but *finding the main idea* was reported the most frequently. This skill was also reported consistently across all cultural groups. In contrast, *summarizing a passage* was given a low rating and was not prevalent in the preparation of students from Germany and Saudi Arabia.

Table 26 presents data on student preparation for the listening portion of the exam and shows that students reported using all the suggested skills. *Finding the main idea* received the highest rating, indicating that student beliefs regarding the listening section are similar to their beliefs about the reading section. There is little variation among mean ratings, which shows that students agreed that all skills were used.

Table 27 shows student responses about writing.

Table 27 presents data on student preparation for the writing section of the exam. Students reported using all of the skills they had been taught, but *organizing a cohesive essay* received the highest mean rating, which indicates that it was the most frequently reported skill. Spelling and vocabulary were also highly rated, which suggests that students see these as important components of writing.

Table 28 shows how often students reported using skills to respond to tasks on the speaking section.

The data in Table 28 show that students most frequently reported using the strategy *express opinions on topics*. *Participate in speech similar to an academic discussion* was the least frequently reported strategy. In addition to not meeting the 2.75 threshold for agreement, this result is consistent with other student beliefs about the speaking section; namely, that students' survey responses indicate a disconnect between the speaking section of the TOEFL iBT and academic language and students' actual speaking ability; Table 20 shows that students are neutral or disagree that the speaking section shows their ability to speak English. This is a common theme in this report.

Summary: Research Question 2

The results for Research Question 2 provide a picture of how instructors and students define academic language ability and to what extent these definitions match the construct as defined on the TOEFL iBT. Within each domain, instructors and students reported a variety of functions as relevant to the university classroom and to the test. Instructors believe

Table 28 Skill Use on the Speaking Section of the TOEFL iBT

On the speaking section of the TOEFL iBT, how often did you use the following skills?	*N* (no response)				*Mean* rating[a]				*SD*			
	G	SA	SK	Total	G	SA	SK	All	G	SA	SK	All
Express opinions on topics	275 (4)	265 (9)	231 (4)	771	3.24	3.27	3.39	3.30	.83	.74	.71	.76
Summarize information verbally	273 (6)	265 (9)	227 (8)	765	3.41	3.14	3.16	3.24	.73	.85	.78	.79
Respond to questions in a timely manner	275 (4)	268 (6)	228 (7)	771	3.39	3.06	2.96	3.14	.81	.88	.82	.84
Talk about thoughts in an organized way	271 (8)	267 (7)	227 (8)	765	3.03	3.00	2.82	2.95	.80	.83	.80	.81
Participate in speech similar to an academic discussion	266 (13)	257 (17)	230 (5)	753	2.56	2.82	2.53	2.64	.98	.90	.90	.93

Note. G = Germany; SA = Saudi Arabia; SK = South Korea.
[a]Minimum = 1; Maximum = 4.

Table 29 Perception of Difficulty of the TOEFL iBT

Country	*N* (no response)	Currently studying in an ESU		Not currently studying in an ESU		Total	
		Mean [a]	*SD*	*Mean*	*SD*	*Mean*	*SD*
Germany	279 (0)	2.86	.472	2.74	.510	2.78	.50
Saudi Arabia	268 (6)	2.32	.466	2.45	.485	2.37	.48
South Korea	235 (0)	2.18	.558	2.05	.5660	2.11	.57

Note. ESU = English-speaking university.
[a]Minimum = 1; maximum = 4.

a variety of academic tasks are important in preparing students for the TOEFL iBT, indicating that the test represents a complex construct of academic language ability within each language domain. These results are further supported by the finding that instructors do not necessarily agree that specific test preparation for the speaking section, including formulaic responses and timed practice, are relevant for preparing students. From the students' point of view, the survey results also show that the test represents a complex notion of academic language ability and that students believe a number of functions are relevant to their performance within each domain. At the same time, concerns remain about the speaking section.

Research Question 3: Do Student Beliefs Differ Based on Context?

The third research question focused on how student beliefs about the TOEFL iBT differ between cultural and educational contexts. To respond to this question, (a) an analysis of student beliefs about the difficulty levels of parts of the test was conducted and (b) possible comparisons between cultural groups (German, Korean, or Saudi) and English language-background (studying in an English-speaking country at present or not) were made. An analysis of the data presented here attempts to determine if these differences exist.

Perception of test difficulty is one way to explore student beliefs about the TOEFL iBT's ability to measure student academic language ability. Table 29 shows perception of difficulty across all survey questions on this topic. Table 29 shows mean ratings by country (Germany, Saudi Arabia, or South Korea) and English-language background (studying at an English-speaking university at present or not).

Mean ratings of the difficulty of the TOEFL were calculated on a 1–4 scale from difficult to easy, respectively. With a mean rating close to 3.0 for both those studying at an English-speaking university and those who are not, this data strongly suggests that German students think the TOEFL iBT is somewhat easy. However, Korean students in both contexts have mean difficulty ratings closer to 2.0 and tend to find the test to be difficult. In both German and Korean groups, those students not currently studying in an English-speaking university think the TOEFL iBT is more difficult than those who are studying in an English-speaking university. The figures are reversed for Saudi students, so that those studying at an English-speaking university find it more difficult than those who are not.

A two-way analysis of variance (ANOVA) was conducted to determine whether cultural group and English-language background affect students' perception of difficulty of the TOEFL iBT. The results showed a significant main effect of cultural group with a small effect size, $F(2, 776) = 107.08$, $p < .001$, $\eta^2 = .04$. There was no main effect of English-language background, $F(1, 776) = 1.03$, $p > .005$, $\eta^2 = .001$. The interaction effect between cultural group and English-language

Table 30 Perception of Efficacy of TOEFL iBT Items by German, Saudi, and Korean Students

		Currently studying in an ESU		Currently not studying in an ESU		Total	
Country	*N* (no response)	*M*[a]	*SD*	*M*	*SD*	*M*	*SD*
Germany	279 (0)	2.96	.42	2.82	.45	2.87	.45
Saudi Arabia	268 (6)	2.58	.55	2.67	.59	2.61	.56
South Korea	235 (0)	2.73	.40	2.75	.45	2.74	.43

Note. ESU = English-speaking university.
[a]Minimum = 1; maximum = 4.

background was not significant, $F(2, 776) = 4.71$, $p > .005$, $\eta^2 = .01$. A post-hoc Tukey test showed that there was a significant difference in beliefs about difficulty between all cultural groups. German students had the least difficulty with the TOEFL iBT, while students from Korea had the most difficulty.

The survey next examined student beliefs about the TOEFL iBT's efficacy and whether they differed by cultural group and English-language background. Table 30 shows the mean results of student beliefs by cultural group and current place of study.

Table 30 shows the differences in means between students from the three cultures and some differences within the cultures. A two-way ANOVA of these data shows that, similar to the previous finding for perception of difficulty, a significant main effect was found for cultural group with a small effect size, $F(2, 776) = 18.01$, $p < .001$, $\eta^2 = .04$, but there was no main effect of English-language background, $F(1, 776) = .03$, $p > .005$, $\eta^2 = .0001$. There was no significant interaction effect between cultural group and language background, $F(2, 776) = 3.47$, $p > .005$, $\eta^2 = .01$.

A post-hoc Tukey test indicates significant differences in means of reported beliefs of the TOEFL iBT's efficacy between all cultural groups. The results of the analysis show that German students think the TOEFL iBT is most efficacious, followed by Korean students, followed by Saudi students. The analysis suggests that cultural group is more important in student beliefs about the TOEFL iBT efficacy than English-language background.

Summary: Research Question 3

Research Question 3 examined whether differences existed between students' beliefs about the TOEFL iBT, and the results reveal that country of origin or cultural background, on this survey, is consistent with significant differences in beliefs. In summary, students from Germany thought the TOEFL iBT was least difficult and most effective. Students from South Korea thought the TOEFL iBT was most difficult and least effective. Interestingly, students from Saudi Arabia studying in an English-speaking university thought the TOEFL iBT was less effective than those not studying in an English-speaking university.

Research Question 4: How Well Do Users Believe Integrated TOEFL iBT Tasks Measure Students' Academic Ability? Do These Beliefs Differ By Context?

The fourth research question has two parts. The first addresses how well users believe that the integrated tasks measure students' academic ability, and the second asks if these beliefs differ by context. Both the student and instructor surveys contained questions designed to address this question. The initial focus groups had revealed that administrators were not sufficiently familiar with the content of the TOEFL iBT to make such judgments; therefore, their survey did not include such questions. Additionally, data from the stimulated recalls addresses these topics. However, the methods and results of the stimulated recalls are discussed in a separate section.

Table 31 explores student and instructor beliefs about the integrated writing and speaking tasks by presenting mean response data from students and instructors, consistent with the 2.75 threshold established. The ratings from both the instructors and students clear this threshold, demonstrating agreement between the two groups that the integrated writing and speaking tasks are a measure of academic language ability.

A t-test shows no significant difference in beliefs about the TOEFL iBT's integrated tasks between instructors and students, $t(556.8) = .06$, $p > .005$, $d = -.003$ for integrated writing tasks and $t(569.59) = -.87$, $p > .005$, $d = .06$ for integrated speaking tasks. Additional analyses were conducted to determine if there were differences between student cultural groups

Table 31 Perception of Integrated TOEFL iBT Writing and Speaking Tasks as a Measure of Academic Language Ability Reported by Students and Instructors

	Integrated writing tasks		Integrated speaking tasks	
User groups	*M*[a]	*SD*	*M*	*SD*
Instructor	3.17 ($n = 266$)	.45	3.10 ($n = 266$)	.45
Student	3.17 ($n = 772$)	.55	3.07 ($n = 769$)	.56

[a]Minimum = 1; Maximum = 4.

Table 32 Efficacy of Integrated Writing Tasks on the TOEFL iBT as a Measure of Academic Language Ability Reported by German, Saudi, and Korean Students

		Currently studying in an ESU		Currently not studying in an ESU		Total	
Country	*N* (no response)	*M*[a]	*SD*	*M*	*SD*	*M*	*SD*
Germany	279 (0)	3.23	.54	3.230	.54	3.23	.53
Saudi Arabia	266 (8)	3.08	.58	3.13	.60	3.10	.59
South Korea	235 (0)	3.28	.50	3.12	.50	3.20	.51

Note. ESU = English-speaking university.
[a]Minimum = 1; maximum = 4.

Table 33 Efficacy of Integrated Speaking Tasks on the TOEFL iBT as a Measure of Academic Language Ability Reported by German, Saudi, and Korean Students

		Currently studying in an ESU		Currently not studying in an ESU		Total	
Country	*N* (no response)	*M*[a]	*SD*	*M*	*SD*	*M*	*SD*
Germany	279 (0)	3.15	.55	3.08	.52	3.10	.53
Saudi Arabia	268 (6)	3.02	.59	3.08	.66	3.04	.61
South Korea	235 (0)	3.120	.51	2.99	.54	3.084	.53

Note. ESU = English-speaking university.
[a]Minimum = 1; maximum = 4.

on beliefs about the integrated speaking and writing tasks. Table 32 shows the means for students in each cultural group and English-language setting.

As depicted in Table 32, there are differences between the students' cultural groups and current setting. Results from a two-way ANOVA show no significant main effect for cultural group, $F(2, 766) = 3.50$, $p > .05$, $\eta^2 = .01$, or for current setting, meaning English-speaking university or not, $F(1, 766) = .84$, $p > .005$, $\eta^2 = .001$. There is also no significant interaction effect between culture group and English language-background, $F(2, 766) = 2.27$, $p > .005$, $\eta^2 = .01$.

Next, examinee beliefs about the speaking tasks are examined. Table 33 shows the means for each cultural group in each setting.

Table 33 indicates that all students agree that the integrated speaking tasks on the TOEFL iBT are effective measures of academic language ability, because all means are above 2.75. German and Korean students at English-speaking universities believed that the TOEFL iBT speaking tasks were slightly more efficacious than those still in their home country, while the Saudi students presented a reverse pattern.

A two-way ANOVA of the data shows no main effect of cultural group, $F(2, 763) = .81$, $p > .005$, $\eta^2 = .002$, and English-language background, $F(1, 763) = 2.79$, $p > .005$, $\eta^2 = .003$, and no significant interaction effect between cultural group and English-language background, $F(2, 763) = 3.37$, $p > .05$, $\eta^2 = .01$, for student beliefs about the efficacy of the speaking tasks.

Summary: Research Question 4

Research Question 4 examined the extent to which students and instructors believe that integrated TOEFL iBT tasks measure students' academic ability and examined if student beliefs differed based on cultural or educational contexts. Survey

Table 34 The TOEFL iBT as an Indicator of International Student Performance

	Mean[a] opinion	*SD*
Administrators	2.90	.61
Instructors	2.65	.72

[a]Minimum = 1; maximum = 4.

results indicate that students and instructors tend to agree that both integrated writing and speaking tasks are a good measure of academic language ability. The beliefs of students and instructors were not significantly different. Additionally, no significant differences were found based on culture or whether students were currently studying at an English-speaking university. Effect sizes for these survey questions were generally small.

Research Question 5: Do Users Believe That Student Performance on Integrated TOEFL iBT Tasks Reflects Performance in Academic Classes? Do These Beliefs Vary by Context?

Similar to Research Question 4, Research Question 5 has two parts: what do users believe and is there difference in beliefs based on context? The survey asked questions of administrators, instructors, and students to determine this. First, administrators and instructors were asked to respond to the statement, "The TOEFL iBT is a good indicator of how international students will perform at my institution." Because administrators did not show familiarity with specific aspects of the test, administrators were only asked this general question. Because it relates to performance in the university, it is included under this research question, to reflect ability on academic classes. The results are shown in Table 34.

Table 34 shows administrators' and instructors' opinions on the strength of the TOEFL as an indicator of international students' performance and success at their institutions. Administrators and instructors differed in their opinions; instructors, overall, were neutral or disagreed with this statement and administrators did agree with it. This was confirmed by a t-test, showing a significant difference with a small effect size: $t(422.7) = -3.8$, $p < .005$, $d = .38$.

Another indication of beliefs about the test and academic ability was gathered via a direct question relating the perception of writing and speaking tasks to student performance in academic tasks. Table 35 displays the perception of integrated writing and speaking tasks as reflecting academic performance by students and instructors.

Table 35 shows that, according to the predetermined threshold, instructors and students agreed that the integrated tasks reflected performance in academic classes for writing and speaking tasks. A *t*-test indicated no significant difference between student and instructor beliefs about whether the writing, $t(1017) = .85$, $p > .005$, $d = .07$, and speaking tasks, $t(1023) = .9$, $p > .005$, $d = .03$, reflect student performance in academic language classes.

The next section examines differences in beliefs across the three cultural groups on the writing and speaking tasks. First, Table 36 presents the means across the three cultural groups in the two English-language backgrounds.

Table 36 shows that all three cultural groups believed that the integrated TOEFL iBT writing tasks reflected academic performance of students. Except for the Saudi student group, all students seemed to believe the tasks are more reflective of student academic performance if they were studying in an ESU. However, a two-way ANOVA indicates no significant interaction effect for cultural group and background, $F(2, 766) = 4.79$, $p > .005$, $\eta^2 = .01$, for student beliefs about the integrated writing tasks. There were no main effects of cultural group, $F(2, 766) = .64$, $p > .005$, $\eta^2 = .002$, and English-language background, $F(1, 766) = .54$, $p > .005$, $\eta^2 < .000$.

Table 35 Perception of Integrated TOEFL iBT Writing and Speaking Tasks as Reflecting Performance in Academic Classes Reported by Students and Instructors

	Integrated writing tasks		Integrated speaking tasks	
User groups	*M*[a]	*SD*	*M*	*SD*
Instructor	2.96 ($n = 247$)	.72	2.90 ($n = 256$)	.80
Student	3.01 ($n = 772$)	.77	2.95 ($n = 769$)	.76

[a]Minimum = 1; maximum = 4.

Table 36 Perception of Integrated TOEFL iBT Writing and Speaking Tasks as Reflecting Performance in Academic Classes Reported by German, Saudi, and Korean Students

		Currently studying in an ESU		Currently not studying in an ESU		Total	
Country	*N* (no response)	*M*[a]	*SD*	*M*	*SD*	*M*	*SD*
Germany	275 (4)	3.04	.83	2.93	.79	2.96	.80
Saudi Arabia	266 (8)	2.95	.81	3.15	.71	3.02	.79
South Korea	235 (0)	3.17	.68	2.94	.70	3.05	.70

Note. ESU = English-speaking university.
[a]Minimum = 1; maximum = 4.

Summary: Research Question 5

Research Question 5 examined the extent to which students, instructors, and administrators believe that integrated TOEFL iBT tasks reflect students' performance in academic classes. However, the focus groups indicated that administrators were not familiar with integrated tasks; therefore, their surveys questions did not ask about the integrated TOEFL iBT tasks. Administrators and instructors did have significantly different beliefs about the test in general, and administrators tended to agree that performance on the TOEFL iBT is a good predictor of performance in academic classes, while instructors tended to disagree. While instructors tended to disagree about many statements about the test in general, they tended to agree that performances integrated tasks reflect student performances in the real world. Student beliefs about integrated tasks were similar to instructor beliefs and also tended to be positive. There were no significant differences between students based on culture or context of study.

Stimulated Recall

Stimulated recall allows the test taker to interact with and respond to a task or tasks and then respond to questions about the experience after the task has been completed (Gass & Mackey, 2000). The purpose of conducting a stimulated recall rather than a think-aloud is that the test taker can complete the test tasks or items in an authentic manner and then reflect on the cognitive processes used to complete the test tasks without being interrupted to think aloud. To facilitate reflection, the researcher can show the test taker the task and his or her response. Such an approach can provide important information to test developers about the cognitive processes test takers use, as well as test takers' beliefs about what the test task is actually measuring.

For this study, a total of 12 stimulated recalls were conducted to provide qualitative information about students' perceptions of TOEFL iBT items. ETS provided a computer program with TOEFL iBT tasks. Researchers determined that the stimulated recalls should last no more than 2 hours and, thus, reviewed the tasks to select those most relevant to the research questions and most prevalent based on student responses during focus groups. While reading and listening tasks were included in this section of the study, the stimulated recalls primarily focused on integrated writing and speaking tasks (Table 37).

Researchers developed an administration and interview protocol for the stimulated recalls. Bilingual consultants were then recruited and trained to administer the recalls and translated the protocol into Arabic, German, and Korean. Each consultant piloted the translated protocol with one student before it was made operational. Appendix G includes an excerpt from the stimulated recall protocol.

Table 37 TOEFL iBT Tasks Used for Stimulated Recalls

Section	Task type
1	Reading
2	Listening
3	Writing (integrated)
4	Speaking (integrated)

Participants

Students were recruited from English-language schools and universities throughout the Washington, DC, metro area. Students were paid $75 for participating in the stimulated recall. The project planned to include two ESL and two university students per cultural group, but difficulty recruiting students led to several changes in the research design. Because of difficulty recruiting Saudi students, Arabic-speaking students from other countries were recruited for the study. For the four participants in each of the Arabic and Korean groups, two were ESL students and two were matriculated university students. However, the four German participants were all matriculated students at an American university.

Stimulated recalls in German were conducted in English rather than German because a German-speaking researcher was not available at the time. German participants reported being comfortable answering questions in English and tended to have high levels of English speaking proficiency.

Despite these changes to the research design, the data were able to adequately address the project's research questions through rich qualitative analysis.

Research Question

This part of the study yielded supplemental data that addressed the fourth research question: How well do users believe integrated TOEFL iBT tasks measure students' academic language ability? Do those beliefs vary by context? Since this part of the study only focused on test takers, the research question could only be answered with regard to the extent to which test takers believe that the integrated tasks measure student language ability. The survey elicited user beliefs on a large scale; the stimulated recalls elicited more fine-grained user beliefs within the immediate context of taking the test.

Method

The purpose of this stimulated recall was to gather feedback from students about test items on the TOEFL iBT. During each stimulated recall session, students completed sections that included four test items from the TOEFL iBT. After completing each section, they were asked questions about the section. For the integrated writing task, students partially completed the task because of time constraints. For the Arabic and Korean groups, a native speaker consultant was trained to conduct the stimulated recall so that participants could respond in their native language or in English. A native English speaker conducted stimulated recalls with German participants. At the conclusion of the interview, students discussed their global views of the test items. All interviews were audio recorded. Table 38 shows the steps taken for each stimulated recall.

Each stimulated recall followed the structure represented in Table 38. The procedures were identical across languages, thus allowing comparison of results.

Results

Because it provides information about students' self-reported perceptions of test items, the stimulated recall data extended the survey results and helped illuminate potential differences between cultural groups. While test scores and constructed responses show how the test format and task characteristics may affect student performance, introspective data provides an explanation of how students perceive the tasks and the factors they think affect their performances. The next sections detail the themes that arose from the data.

Coding and Analysis

After each stimulated recall, the consultant transcribed the interview recording. Arabic and Korean transcripts were later translated into English. As stated above, interviews with German students were conducted in English. Each consultant also submitted a short report on topics that emerged during the interview and any issues that seemed especially important to the student. The reports were analyzed by two researchers to establish common themes for coding across all interviews. Table 39 shows the themes that emerged for coding of the transcripts.

The coding categories in Table 39 were then applied to the translated transcripts. In the interview protocols, students were explicitly asked to discuss the difficulty and efficacy of the tasks, and subthemes emerged across several respondents

Table 38 Steps of Protocol Followed for Stimulated Recalls

Step	Description	Responsible party	Comments
Step 1: Background paperwork	Complete background paperwork, including waiver	CAL staff	
Step 2: Audio recording	Turn on audio recorder	CAL staff	
Step 3: Introduction	Introduce project	Consultant (target language) CAL staff (English)	Read from script
Step 4: Reading	Administer Section 1: Reading	CAL staff	Consultant takes notes using observation protocol
Step 5: Ask questions about reading	Ask questions about Reading Section	Consultant	Read from script
Step 6: Listening	Administer Section 2: Listening	CAL staff	Consultant takes notes using the protocol
Step 7: Ask questions about listening	Ask questions about the Listening Section	Consultant	Read from script
Step 8: BREAK		Student breaks for 5 – 10 minutes	
Step 9: Writing	Administer Section 3: Writing	CAL staff	Consultant takes notes using the protocol
Step 10: Ask questions about writing	Ask questions about the Writing Section	Consultant	Read from script
Step 11: Speaking	Administer Section 4: Speaking	CAL staff	Consultant takes notes using the protocol
Step 12: Ask questions about speaking	Ask questions about the Speaking Section	Consultant	Read from script
Step 13: General questions	Ask general questions about test	Consultant	Read from script
Step 14: Wrap-up	Conclude interview	Consultant	Read from script
Step 15: Payment	Pay test taker	CAL staff	

Note. CAL = Center for Applied Linguistics.

Table 39 Themes for Coding Stimulated Recalls

General theme	Subthemes
Difficulty	Difficulty of specific task
Task reliance on a specific skill	Listening, reading, speaking, writing
Fluency	Extent to which task required fluency
Administration method	Test format/computer delivery
Expected Score	Extent to which the student felt s/he would receive the expected score
Task content	How task reflected academic content Task not effective/does not match academic tasks
Task effectiveness	Relevance of task to student's academic field
Student background	
Time issues	
Student understanding of prompt	Prompt comprehended/not comprehended

that were particularly relevant to these topics. All transcripts were independently coded by two researchers, and any discrepancies were discussed with a third researcher until consensus was reached. Although the stimulated recalls addressed all four skills, speaking and writing tasks were particularly important for participants. The major themes that emerged across all contexts were the challenges test takers saw in these tasks, as well as their reflections on the integrated writing tasks. The coding process was used to identify prominent issues in the data. Results are presented here descriptively by theme.

Challenges of Integrated Speaking Tasks

Several themes emerged across participants. Each student noted the difficulty of the speaking section, and most students rated it as the most difficult section of the TOEFL iBT. The challenge of the speaking section for students supported the survey data, for which student average ratings were lower for speaking than for any other skill. Two factors emerged as the perceived cause of these difficulties. For students with lower scores on the TOEFL iBT, the skill of speaking itself was

perceived as being more difficult than the other skills on the test. This topic was most frequently mentioned by Korean students.

In contrast to the focus groups, students did not mention the use of formulaic responses or patterns. Focus group data showed that a rehearsed response (a kind of formulaic speech or memorized response pattern) that candidates apply to the speaking responses was a key test preparation strategy for many students, especially among Korean students. That this strategy was not mentioned in relation to actual TOEFL iBT tasks may reveal a difference between beliefs about test preparation and the skills actually needed to successfully complete the tasks. When discussing their performance on the integrated speaking tasks, students mentioned the high cognitive load required by the task content (e.g., understanding and summarizing the reading in order to formulate a response). It may also be the case that on the actual tasks, the cognitive demands of the content require enough engagement from students that formulaic response patterns are not relied upon. Recall data supports this, in that most students expressed concerns about the content of their answers. One student said:

> I think it's more about summarizing and thinking what I read in the different sections. And not all about the English I talked. I had kind of problems to remember what was in first section and what was in second section and I read this sentence, but I'm not sure if it was still where.

Students perceived the test format as another source of task difficulty. Many students mentioned the difficulty of speaking to a computer rather than a human. Several students mentioned that the computer format is difficult because the computer provides no response or feedback. One Korean student commented, "You can't actually have a person sit in front of you and practice this section at all. It takes time for the person and it's really hard to get feedback."

Additionally, the computer-based format includes timed planning and responses. All students mentioned this issue in relation to task difficulty. Although in general students seemed to think the planning and response time was too short, there was not a consistent pattern across students. For some questions, students reported that the response time was too long. This difference seems to be due to individual differences, rather than a pattern across language groups or proficiency level. Below, an extended response from a German student's interview illustrates several of these issues (this interview was conducted in English):

> G05: For me, it's definitely the hardest section of the whole test. And I can't really figure out why. It might be due to the fact that I'm a little bit slow. And I don't talk very fast and I always, I need a lot of time to make a point, like, I think maybe I would talk too long around the topic without making a real point. I mean, now, it was easier for me than when I took the test in Germany because here I'm more used to talk English. But then it was just like, I couldn't do it. For every single question I didn't get to say everything I wanted to say. Because the time amount is very short, like the preparation time is short, and also like the time you can actually talk is not a lot.
>
> Interviewer: So do you think that it shows how well you can speak in English?
>
> G05: Well, yeah, it does, but it's very hard to focus and to concentrate on the question in this short time, and I also think it's more difficult if you have to speak into a microphone than like to have an actual interview with another person. That makes it difficult.

In some cases, students expressed the belief that time pressure negatively affected their performance and scores. Most students also described the timed format as inauthentic. However, it should be noted that overall the students felt the type of tasks included in the TOEFL iBT reflect the type of speaking activities present at the university level. Format, rather than task function, caused the perceived difficulty.

Although students did not mention the use of formulaic or memorized responses in the integrated speaking tasks, many students mentioned the importance of structured or organized responses and one student discussed using a response template. For example, students mentioned the need to produce "short organized answers" and to "structure your talk" on the tasks. The use of notes was mentioned in relation to organizing answers, and several students described using their notes to make an organizational plan for their speech. One student mentioned using his notes to write down key words from the listening passage to include in his speaking response. The students' descriptions of organized responses reveal the type of language they produce on the test and the cognitive processes they use to complete the integrated speaking

tasks. Students reported that they relied on multiple skills when they listen to a passage, take notes, and then answer the speaking prompt. However, students also reported that they use formal and structured modes of oral language. This finding is not surprising and is consistent with other research on technology-mediated oral language testing.

Challenges of Integrated Writing Tasks

On integrated writing tasks, students are required to read a text, listen to a lecture, and write an essay about the passage and lecture. For the purposes of the stimulated recall, students did not complete the entire writing task and were stopped after a limited amount of time to provide their opinion about the tasks. Participants then reviewed the task and their writing in order to answer questions about their perceptions of and performance on the task. Questions included information about the task purpose, the skills needed to complete it, differences and similarities between the test and real-world writing tasks, and perceived task difficulty.

When asked to describe the skills needed to complete the integrated writing task, all students mentioned the importance of summarizing information. When reading and listening, students reported that they read or listened for key points and made notes about these points to use in their essays. Many students mentioned that they knew both passages would contain three key points and the purpose of their reading or listening was to identify these points. Note taking was a central activity for every participant in the study. Students perceived the purpose of the writing tasks as summarizing the texts and discussing any differences of opinion between the reading and listening. Although all students agreed that the task required summarizing skills, about half of the students mentioned that the task did not allow them to use analytical skills or to express their own opinion about the topic. As one student stated, "Now I had assignments which were more analytical and more our own opinion. So it's not just reproducing what someone said or which stance you were."

When describing differences between the test task and university-level tasks, the use of analytical skills was mentioned by several students as a key difference between the test and real-life writing. At the university level, students argued, writing tasks go beyond summarizing. This response was especially common among the graduate-level participants, who frequently mentioned the need to support and defend a position in academic writing. However, a few students mentioned that the summarizing skills used in the integrated task are similar to the skills needed to write literature reviews in academic essays. As one student said in response to a question about whether the writing task was similar to real writing tasks in U.S. universities, "Yes, definitely, yes. It's kind of similar. But of course you can't organize it so well in 20 minutes and that would do for paper."

Consistent with the comment about time, students also perceived differences in terms of the test tasks and real-life tasks in terms of the time limits imposed on the writing test. Many students described real-life writing tasks as ongoing rather than immediate. Writing involves extended thought and research, and many students expressed the opinion that the timed nature of the test was somewhat inauthentic and may have a negative impact on their performance. Interestingly, one Korean student felt that his performance on the test was better than his actual ability. He attributed this difference to the fact that he had practiced extensively for the exam and scored well because he was able to use a response template.

As with the integrated speaking tasks, focus group data suggested that students practiced formatted responses for the writing section, with Korean students frequently mentioning this strategy. This topic came up with two students during the stimulated recalls; both students were from South Korea. These two students mentioned that during the reading and listening section they were prepared to identify three main points. They also expected the listening passage to oppose the perspective articulated in the reading passage. One student mentioned using the main points from the reading passage helped him predict the information that would be provided in the listening section. Finally, these students mentioned using a template to formulate their response. In general, all four of the Korean participants discussed test-taking strategies more frequently than other participants. One student also mentioned that in Korea the test questions are often known beforehand and circulated among students. This finding came up in the focus groups as well. Arabic- and German-speaking participants did not mention practicing for writing tasks or the use of specific templates.

Although participants believed that writing for the purpose of summarizing was not an authentic academic task, students in general felt that the integrated writing task was a good measure of their academic writing ability. Perceptions of task difficulty varied across participants and did not follow a discernible pattern based on cultural context but, rather, based on English proficiency. High proficiency students perceived the task as being easier than lower proficiency students did, with the exception of the one student who felt that he performed better on the test than his actual ability. Students without experience in academic classes in U.S. university contexts perceived the most difference between the integrated

writing tasks and actual academic tasks, and several participants discussed cultural differences between university study in their native countries and the United States. For example, two of the German students mentioned that, in Germany, writing tasks are usually large projects that occur at the end of a course of study, such as an undergraduate thesis. Small writing assignments such as the one presented on the test were not part of their educational experience. However, many of the participants mentioned that they either expected to experience or had experienced frequent writing tasks in the U.S. educational context.

Integrated Tasks

Overall, data from the stimulated recalls show that participants from each cultural background perceive the speaking and writing integrated tasks as being good measures of their actual English abilities and that the tasks are in some ways reflective of the types of tasks that occur in U.S. academic settings. As one student pointed out in response to the question, "Do you think it's important to have some questions or tasks where you have to read and listen then speak, or read this and write?" The student responded:

> Yeah, yeah, I think it helps, because if you have both, you get a better idea of the topic, so it's better for the participants to have both sources. Yeah, it's more helpful than harmful. And I think it's more of the real life than just reading.

For both speaking and writing, task time constraints were a frequently mentioned source of frustration for students and a main factor cited when discussing task inauthenticity. As one student pointed out, "I had to prepare for the answer. There are 15 seconds to think. In that 15 seconds, I have to write down the key words that I am going to say, but just now I wrote the wrong thing."

Summary: Stimulated Recalls

The stimulated recalls produced a great deal of rich information about student beliefs about the TOEFL iBT. Specifically, most students believe that the integrated tasks reflect to some degree the type of tasks they need to perform in university classrooms in the United States; in other words, most students believe that they need to listen and read to academic material and then respond orally or in writing for academic contexts. For almost all students, though, the timed nature of the tasks is the least authentic way that the integrated tasks reflect how students need to use writing and speaking in academic settings.

Conclusions

This study has investigated the beliefs of three groups of stakeholders on the TOEFL iBT: administrators, instructors, and students. The purpose of the study was to investigate what these three groups believe about the TOEFL iBT and how they define academic language ability, as well as whether any differences exist between or within groups about these beliefs.

Analysis of the data suggests that there are differences in beliefs between administrators and instructors, a lack of differences in beliefs between students and instructors, ongoing issues regarding beliefs on the speaking tasks, and differences in beliefs between the student groups. Conclusions are described for each research question.

Research Question 1

What are users' beliefs about what the TOEFL iBT measures? The focus groups, surveys, and stimulated recalls showed that instructors generally agree that the TOEFL iBT can show how well students can read, write, listen, and speak in English, while most students do not agree that the TOEFL iBT measures their ability to speak English. The speaking tasks emerged as the most problematic across all groups and on all measures. The issue of the speaking tasks first emerged in the focus groups, and the survey results show that the speaking tasks received the least favorable responses from students and instructors. Overall, students in all contexts did not agree or were neutral that the speaking section allowed them to demonstrate how well they could speak English. Indeed, Pearlman (2008) points out that the speaking sections (both

integrated and independent) are novel for this type of test and require different preparation on the part of examinees. Stimulated recall data suggest that difficulties with the speaking tasks are related to the format, rather than the content, of the tasks. In examining the emerging data from the focus groups, combining it with the results of the surveys and qualitative data from the stimulated recalls, this study suggests that the way speaking test data are elicited was perceived as problematic by students and instructors and that the TOEFL iBT does not necessarily reflect how students and instructors believe that speaking takes place in authentic academic environments.

Research Question 2

How do users define academic language ability in their contexts? This research question was addressed more directly in the focus groups than in the surveys. In the focus groups, students, instructors, and administrators were able to discuss the extent to which the TOEFL iBT showed how well students know English, and this line of questioning allowed the research team to develop specific questions for the survey. However, such an open-ended question could not be asked overtly of students, administrators, and instructors on the survey, because the very term *academic language ability* is specific jargon in the field of language testing. Instead, students and instructors were asked to report which functions were relevant to the university classroom and the TOEFL iBT. Instructors reported that they believe that a variety of academic tasks are relevant to preparing students for the TOEFL iBT, and this result suggests that the test represents a complex construct of academic language ability within each language domain. These results are further supported by the finding that instructors do not necessarily agree that specific test preparation for the speaking section, including formulaic responses and timed practice, are relevant to preparing students.

While the surveys and focus groups revealed some overall dissatisfaction with the test among instructors, the survey results suggest that this belief is not a result of the test's content. From the students' point of view, the survey results also show that the test represents a complex notion of academic language ability and that students report using a variety of academic skills within each domain to accomplish test tasks, which provides further evidence for the complexity and real-life nature of the tasks.

Research Question 3

Do student beliefs differ across contexts? This research question was explored in the survey, and it allowed us to focus on the beliefs of students from one country in three major world regions: Germany (Europe), Saudi Arabia (Middle East), and South Korea (Asia). One major theme throughout the qualitative and quantitative measures on the study was whether students had the same or different beliefs about the TOEFL iBT based not only on their country of origin, but also based on whether they were currently studying at an English-speaking university. The results for this part of the research were mixed. First, there were significant differences in beliefs between the student groups. Among the three groups, Korean students ranked the TOEFL iBT with the highest level of difficulty, while German students ranked it with the lowest level of difficulty. These differences were significant, suggesting that culture of origin shapes user beliefs about the TOEFL. However, in cases where cultural differences between groups were statistically significant, effect sizes tended to be small, suggesting that cultural context may have limited practical importance as an explanation for student beliefs. While the survey results reveal limited cultural differences between groups, the qualitative data from the focus groups and stimulated recalls show patterns of difference between contexts in areas such as test preparation and beliefs about integrated speaking tasks.

For students from all cultural contexts, beliefs did not differ based on educational context. As previously mentioned, Stricker and Attali (2010) found that attitudes differed by world area and test section. They found that students from Germany were either neutral or negative toward the TOEFL iBT, while students from China, Columbia, and Egypt held more positive views toward the test. Moreover, Stricker and Attali found that test-taker attitudes were most positive toward the listening and writing sections and least positive toward the speaking sections. The data from this study are incongruous with some of Stricker and Attali's findings; for example, the results of this study showed that German students generally had positive perceptions about the TOEFL iBT. However, this study corroborates Stricker and Attali's findings that students are least positive toward the speaking sections. All aspects of the data collection, from focus groups to surveys to stimulated recalls, indicate that test takers are least positive toward the speaking section.

Research Question 4

How well do users believe that the integrated TOEFL tasks measure students' academic language ability? Do these beliefs differ by context? The survey results showed that both students and instructors agreed that both the integrated speaking and writing tasks were effective in measuring both student academic language ability. There was no significant difference in beliefs between the two groups, both of which were familiar with the content of the TOEFL iBT. This finding suggests that the integrated tasks are well received by the target population. Overall, the results for Research Question 4 provide some of the study's strongest evidence for the validity of the TOEFL iBT and integrated tasks. The match between the test construct and users' perceptions supports the use of the test for university admissions. Additionally, the similarities between instructor and student beliefs indicate that information about the test is being disseminated consistently and communicated effectively to different user groups. Although instructors and students need different information about the test and use this information for different purposes, the test construct should be clear to both groups.

Research Question 5

Do users believe that student performance on integrated TOEFL iBT tasks reflects performance in academic classes? Do these beliefs vary by context? Administrator beliefs about integrated tasks were not analyzed because this group had limited exposure to test content. In looking at the test in general, the survey results showed that administrators and instructors differ in their beliefs about whether the TOEFL iBT is a good indicator of student performance in academic classes. Instructors' mean rating of belief indicated that they did not meet the determined threshold for agreement with this statement, while administrators' results showed that they did. There are a few explanations for this difference. In discussing the administrators' versus instructors' views, it was suggested that administrators' beliefs about the TOEFL iBT as a good predictor of student performance may be due to administrators' broad perspective on international students' performance across the university (S. Ross, personal communication, October, 2009). In other words, administrators have access to student records, including admission and grades, while instructors have more limited access to admission records and often only to grades for a small group of students. At the same time, it is possible that administrators have a wider view of student grades, graduation rate, and other factors that may lead to their success at an English-speaking university, while instructors' view may be more limited to their specific performance in English-language classes. Therefore, the difference between instructor and administrator beliefs about the TOEFL iBT's predictive validity may actually support the TOEFL iBT's validity argument, because administrators are able to take a wide view of the TOEFL as an admission tool, as well as its relationship to future student performance.

In examining responses to questions about the integrated speaking and writing tasks in particular, the survey showed that both students and instructors tended to agree that these tasks reflect performance in academic classes. This result highlights an interesting discrepancy between how instructors perceive the test as a whole and their beliefs about integrated tasks. In other words, although instructors did not agree that the TOEFL iBT is an accurate predictor of how well a nonnative English speaker will perform at an English-speaking university, when presented with a sample integrated task on the survey, instructors did tend to agree that integrated tasks reflected performance in academic classes.

Limitations of the Study

This study was designed to investigate user beliefs about the TOEFL iBT through multiple methods: focus groups, surveys, and stimulated recalls. Throughout each part of the study, challenges in recruitment and participation rates led to some limitations in the results. For example, it was not possible to fill all focus groups, and it was necessary to recruit students from other Arabic-speaking countries and not limit the focus groups to students from Saudi Arabia. This was also true for the stimulated recalls. Additionally, although the response rate for the student sample, based on e-mail addresses provided by ETS, was 17%, slightly lower than the expected response rate of 20%, the sample size for the survey exceeded the target number. It is also important to note that instructors and administrators were recruited via direct e-mail messages, LISTSERVs, and social networking sites. Therefore, it is impossible not only to determine a response rate, but also to discuss the representativeness of the sample for administrators and instructors except in terms of their self-described backgrounds. The profile of administrators and instructors of the target population is diverse and variable, and all results should be interpreted with these limitations in mind.

It is also important to note a lack of variability in responses to the questions on the survey. This lack of variability may be due to wording of survey questions; future survey research should examine the wording of questions to determine if different results are elicited from similar populations with differently worded questions.

Recommendations

This study sheds light on user beliefs, including administrators', instructors', and students' beliefs, regarding the TOEFL iBT. As it has only been in use for a short time, it is likely that many of these user beliefs continue to be shaped by beliefs about previous forms of the TOEFL. Nonetheless, a few recommendations emerge from both the qualitative and quantitative data.

Recommendation 1: Educate Administrators and Instructors

Focus groups revealed that administrators had a global view of the impact of the TOEFL iBT on student admission and its relationship with academic performance. At the same time, many administrators in focus groups were still better versed in previous scoring systems for the TOEFL and often referred to these scores over the new ones. The administrators who responded to the survey revealed that they were experienced administrators and had worked for a long time with the TOEFL and the students who take it.

While administrators appear to have an overall positive belief system about the TOEFL iBT, they do not necessarily have an accurate idea of how it is different from previous versions of the TOEFL. Educating administrators, who believe more strongly than instructors that the new TOEFL iBT is effective, could result in even more positive beliefs about the TOEFL iBT; it could also help inform changes to the TOEFL. In addition, longitudinal studies that track student progress in university settings may provide valuable information on entrance score requirements and provide additional information for university administrators as they make decisions about students' admission into academic programs.

Similarly, instructors' responses to the survey indicated a lack of awareness about the TOEFL iBT changes. While ETS has worked to convey the content of these changes, it is possible, and, based on the results, likely, that the changes have not been communicated in venues that instructors access and in ways that are accessible and understandable. It might be helpful to review existing materials to describe these changes with instructors to determine whether instructors are aware of these materials and determine whether the message about the changes is being conveyed adequately.

Recommendation 2: Investigate Speaking

Students had the least positive beliefs about the speaking section, as supported by quantitative data from the survey, as well as qualitative data from focus groups and stimulated recalls. Investigating the relationship between the speaking tasks on the TOEFL and the language needed for success in academic classes in English is an important first step. In addition, conveying these results and any additional steps that will be taken to improve the speaking sections would no doubt be beneficial to the long-term success of the program.

However, it is important to note that speaking is a difficult domain to test authentically and even more difficult to test on a large scale. It may be helpful to survey administrator, instructor, and student attitudes about the ability of any speaking test to capture academic speaking; indeed, users may believe that no test can adequately test speaking. A comparison between user beliefs about testing speaking and user beliefs about the TOEFL iBT's efficacy in testing speaking may provide helpful results.

Recommendation 3: Consider Examining Test Preparation Strategies by Context

Test preparation is an important business for the TOEFL program; at the same time, students have mixed beliefs about how to best prepare for the TOEFL. It may be that test preparation needs to be conducted different ways for students from different cultural backgrounds. Test preparation is particularly relevant for the writing and speaking portions of the TOEFL iBT investigated in this study. Students' beliefs about how to prepare for and perform well on the test indicate that

students favor structured practice and, in some cases, rely on prepared answer formats. Examining the impact of the test on student learning may lead to further improvements in both the test and test preparation activities. In addition, beliefs that the TOEFL iBT is more appropriate and difficult for some cultural groups than for others would be an interesting belief to investigate further, in order to provide evidence for test fairness across groups.

Recommendation 4: Conduct Longitudinal Studies With Administrators

In addition to investigating the efficacy of TOEFL iBT tasks from a test development perspective, it may also be important to conduct studies with administrators over time. As mentioned earlier, administrators have fairly positive overall beliefs about the TOEFL iBT; it would be helpful to determine if these beliefs change as the TOEFL iBT becomes more widely used. In particular, the longitudinal study might investigate uses of the TOEFL iBT for placement. Administrators rely on test scores for university entrance requirements. In some cases, admitted students may need additional language support services. Administrator knowledge about TOEFL iBT scores, including domain subscores, may provide a fruitful area for further investigation into test use, as well as appropriate and strategic assistance for students in academic and English-language programs.

In general, this study has found much support for the construct of the TOEFL iBT, and most users report that, as designed, students are expected to read and listen to information and then demonstrate their understanding of or opinions toward the material via speaking and writing. Overall, the results of this study generally support the validity argument for the TOEFL iBT. At the same time, more research on differences in contexts among test-taker beliefs, the education of administrators, and changes to the speaking section could help to support the validity argument and improve the test.

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, *13*(3), 280 – 297.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453 – 476.

Brindley, G. (1998). Outcomes-based assessment in language learning programmes: A review of the issues. *Language Testing*, *15*(1), 45 – 85.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples of emerging insights. *Language Testing*, *18*(4), 393 – 407.

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning*, *45*(2), 251 – 281.

Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

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

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, *21*, 107 – 145.

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. MS-26). Princeton, NJ: Educational Testing Service.

Darling-Hammond, L. (1994). *Professional development schools: Schools for developing a profession*. New York, NY: Teachers College Press.

Educational Testing Service. (2007). *Test and score data summary for TOEFL Internet-based test: September 2005 – December 2006 test data Test of English as a Foreign Language.* Retrieved from http://www.ets.org/Media/Research/pdf/TOEFL-SUM-0506-iBT.pdf

Educational Testing Service. (2008). Validity evidence supporting the interpretation and use of TOEFL iBT™ scores. *TOEFL iBT Research Insight Series, 1*(4). Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Educational Testing Service. (2010). *Test and score data summary for TOEFL Internet-based and paper-based tests: January 2009 – December 2009 test data.* Retrieved from http://www.ets.org/Media/Research/pdf/test_score_data_summary_2009.pdf

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, *10*(3), 235 – 254.

Enright, M., & Cline, F. (2002, March). *Evaluating new task types for TOEFL: Relationships between skills*. Paper presented at the annual TESOL Convention, Salt Lake City, UT.

Epp, L., & Stawychny, M. (2001). Using the Canadian Language Benchmarks (CLB) to benchmark college programs/courses and language proficiency tests. *TESL Canada Journal*, *18*(2), 32 – 47.

Freedman, S. (1991). *Evaluating writing: Linking large-scale testing and classroom assessment.* Occasional Paper 27. University of California, Berkeley: Center for the Study of Writing.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Grant, L. (1997). Testing the language proficiency of bilingual teachers: Arizona's Spanish proficiency test. *Language Testing*, *14*(1), 23–46.

Griffin, P. (1995). *The American literacy profile scales: A framework for authentic assessment*. Portsmouth, NH: Heinemann.

Hamp-Lyons, E. (2007). The impact of testing practices on teaching ideologies and alternatives. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching: Vol. 15,3* (pp. 487–504). New York, NY: Springer US.

Hamp-Lyons, E., & Brown, A. (2006). *The effect of changes in the new TOEFL format on the teaching and learning of ESL/EFL: Report of the development and conduct of the baseline study (2002–2005).* Unpublished manuscript.

Hamp-Lyons, L., & Shohamy, E. (2004). *The effect of changes in the new TOEFL format on the teaching and learning of EFL/ESL: Stage 1 (2002–2003): Instrument development and validation.* Unpublished manuscript.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and progetto lingue 2000*. Cambridge, England: University Press.

Hoge, R., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297–313.

Jamieson, J., Eignor, D., Grabe, B., & Kunnan, A. (2008). Frameworks for a new TOEFL. In C. Chapelle, C. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55–95). New York, NY: Routledge.

Jamieson, J., Taylor, C., Kirsch, I., & Eignor, D. (1999). *Design and evaluation of a computer-based TOEFL tutorial* (TOEFL Research Report No. RR-62). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1999.tb01799.x

Kern, R. (1995). Students' and teachers' beliefs about language learning. *Foreign Language Annals*, *28*(1), 71–92.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, CA: Sage.

Linn, R. L., Baker, E. L., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling*. Thousand Oaks, CA: Sage.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal*, *38*(1), 73–95.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, *23*(4), 445–465.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, *19*(1), 33–56.

Pearlman, M. (2008). Finalizing the test blueprint. 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. 227–258). New York, NY: Routledge.

Powers, D., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment*, *1*(2), 153–173.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. MS-21). Princeton, NJ: Educational Testing Service.

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement*, *38*(2), 265–273.

Schmitt, N., Gilliland, S. W., Landis, R. S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology*, *46*(1), 149–165.

Sharpley, C., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools*, *23*(1), 106–111.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, *13*(3), 298–317.

Stansfield, C. W., & Kenyon, D. M. (1996). Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. In A. Cumming, & R. Berwick (Eds.), *Validation in language testing* (pp. 124–153). Clevedon, England: Multilingual Matters.

Stricker, L., & Attali, Y. (2010). *Test takers' attitudes about the TOEFL iBT* (TOEFL iBT Report No. TOEFL iBT-13). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02209.x

Stricker, L., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based Test of English as a Foreign Language. *Computers in Human Behavior*, *20*, 37–54.

Takagi, A. (2010). *A critical analysis of English language entrance examinations at Japanese universities* (Doctoral dissertation). Retrieved from http://hdl.handle.net/10036/117893

Teachers of English to Speakers of Other Languages (TESOL) (1998). *Managing the assessment process: A framework for measuring student attainment of the ESL standards*. Alexandria, VA: Author.

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. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02024.x

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 Report No. TOEFL iBT-05). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02123.x

Appendix A
Focus Group Screening Questionnaire

Students

- Have you taken the Internet-based TOEFL (iBT) in the past 18 months?
- When did you take the iBT?
- Are you enrolled in ESL or university courses?
- Do you have a copy of your score report that you can bring to the focus group? (This is not necessary for participation in the focus group.)
- What was your iBT score? (Ranges)
- Have you taken an Internet-based TOEFL (iBT) preparation course?
- What preparation courses have you taken?

Instructors

- Have you taught TOEFL preparation courses since September 2005 (when the iBT was launched)?
- Do you work with students who are preparing for the Internet-based TOEFL (iBT)?

Appendix B
Selected Focus Group Protocol Questions, Student and Instructor Groups

Table B1 *Group, Topic, and Sample Questions*

Group	Topic	Sample Questions
Students	English courses	• What do you do in a typical iBT preparation class? • How well do you think these classes prepared you to take the iBT? • How well do you think these classes have prepared you to study in the United States?
	TOEFL iBT	• What information do you think the test developers can find out about you from the iBT? • How difficult or easy was the iBT for you? • Did you feel prepared to do well on the iBT? • You don't have to talk about your score, but do you think the score you got shows how well you know English?
Instructors	Teaching English courses	• How do you prepare students for the iBT? • [PROBE] Can you give me an example? • What do you think is the most important thing you can do to prepare students for the iBT? • [PROBE] Can you describe this?

Table B1 Continued

Group	Topic	Sample Questions
	TOEFL iBT	• How many students do you have each year who take the iBT? • How familiar are you with the iBT? • [PROBE]: What kinds of tasks are on the iBT? • What do you think of the tasks on the iBT? • What type of language skills do students need to do the tasks on the iBT? • What information do you think the test developers find out about students from the iBT? • The iBT is different than other versions of the TOEFL because the tasks are integrated. Why do you think the iBT was developed? • Do you think the scores your students have received on the iBT show how well they know English?

Appendix C
Selected Questions: Student Survey

What is the highest level of education you have completed?

- High school/secondary school diploma
- College/undergraduate degree (AB, BA, BS)
- Graduate degree (Master's, Doctoral, etc.)
- Other (please specify)

At which level are you currently studying? (Select all that apply.)

- High school/secondary school
- College/undergraduate level
- Graduate level (Master's, Doctoral, etc.)
- I'm not a student
- Other (please specify)

How long have you studied English?

- 1 year or less
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- More than 6 years

Did you study English in: (Yes/No)

- Kindergarten (age 3 – 6)
- Primary school (age 7 – 11)
- Secondary school (age 12 – 17)
- College/university
- Other (please specify)

In your home country, how often: (Always, Often, Sometimes, Rarely, Never, Not applicable)

- Did your teachers of English speak to you in English?
- Do you read texts in English in your subject area?

In the past, have you studied in a country where English is the primary language?

- Yes
- No

Which English-speaking country did you study in? (selected response)
Which of the following were you enrolled in? (Select all that apply.)

- ESL class(es) at a private language school
- ESL class(es) at a university or college
- Academic class(es) at a university (i.e., Biology, Business, Engineering, etc.)

Are you currently studying in an English-speaking country?

- Yes
- No

Which English-speaking country are you currently studying in?
Which of the following are you enrolled in? (Select all that apply.)

- ESL class(es) at a private language school
- ESL class(es) at a university or college
- Academic class(es) at a university (i.e., Biology, Business, Engineering, etc.)

Why did you take the Internet form (iBT) of the TOEFL? (Select all that apply.)

- It was the only form available.
- I prefer computer tests to other available testing modes.
- My program or university only accepts iBT scores.
- I received my scores faster by taking the Internet-based test.
- The iBT shows my language abilities better than other TOEFL forms.
- I was prepared in school or class to take the Internet-based version only.
- Other (please specify)

How much did you worry about taking the TOEFL iBT?

- Very much
- Some
- Very little
- Not at all

Have you taken tests similar to the TOEFL iBT? (Select all that apply.)

- International English Language Testing System (IELTS)
- London Tests of English
- Michigan English Language Institute College Entrance Test (MELICET)
- STEP – The Eiken Test in Practical English Proficiency
- Test of English as a Foreign Language paper-based test (TOEFL pBT)
- Test of English for International Communication (TOEIC)
- Test of Spoken English (TSE)
- I did not take a similar test.
- Others (please specify)

Did you prepare for the TOEFL iBT?

- Yes
- No

For how many months did you prepare for the TOEFL iBT?

- 1 month or less
- 2 Months
- 3 Months
- 4 Months
- 5 Months

- 6 Months or more
- I did not prepare for the test.

What is important to do when preparing to take the TOEFL iBT? (Select all that apply.)

- Practice spelling
- Read academic texts
- Practice pronunciation
- Take a practice TOEFL iBT
- Practice academic vocabulary
- Speak with native English speakers
- Study many different subjects in English
- Take a class specifically for the TOEFL iBT
- Practice the TOEFL iBT question format
- Read ETS tutorial on how to take the TOEFL iBT
- You cannot prepare for the TOEFL iBT
- Other (please specify)

Did you take: (Yes/No)

- a course with "TOEFL" in the title; or
- an English course that had a section on the TOEFL?

How useful were the different skills you practiced in your English class for the TOEFL iBT? (Very useful, Useful, Somewhat useful, Not useful, Didn't do this)

- Reading
- Writing
- Listening
- Speaking
- Grammar
- Vocabulary
- Practice TOEFL iBT tests
- Test-taking skills (outlining main ideas, taking notes, answering multiple-choice questions, etc.)
- Using two or more skills (reading, writing, listening, speaking) to complete a task

The preparation course prepared me: (Strongly agree, Agree, Disagree, Strongly disagree)
Please indicate how difficult the sections on the TOEFL iBT were for you. (Very easy, Easy, Difficult, Very difficult)

- Listening
- Reading
- Writing
- Speaking

Please indicate your level of agreement with the following statements. (Strongly agree, Agree, Disagree, Strongly disagree)

- The test questions on the TOEFL iBT felt natural.
- The writing section on the TOEFL iBT let me show how well I can write in English.
- The speaking section on the TOEFL iBT let me show how well I can speak in English.
- The reading section on the TOEFL iBT let me show how well I can read in English.
- The listening section on the TOEFL iBT let me show how well I can listen in English.
- The integrated questions that combine skills (listening, reading, writing, and speaking) let me show how well I can use English.
- Integrated questions that combine skills (listening, speaking, reading, and writing) show my English ability in real-life situations better than questions that measure one skill at a time.

On the reading section of the TOEFL iBT, how often did you use the following skills: (Often, Some, Rarely, Never, Don't know)

- Find the main idea
- Organize information
- Summarize a passage
- Find the relationships between ideas (i.e., compare/contrast, cause/effect, etc.)

On the listening section of the TOEFL iBT, how often did you use the following skills: (Often, Some, Rarely, Never, Don't know)

- Take notes
- Find the main idea
- Find the speaker's purpose
- Draw conclusions based on what is implied
- Make connections between pieces of information in a conversation or lecture

On the writing section of the TOEFL iBT, how often did you use the following skills: (Often, Some, Rarely, Never, Don't know)

- Use vocabulary accurately
- Organize a cohesive essay
- Follow spelling conventions
- Use a range of grammatical structures
- Use idiomatic expressions appropriately
- Identify one main idea and some supporting points

On the speaking section of the TOEFL iBT, how often did you use the following skills: (Often, Some, Rarely, Never, Don't know)

- Express opinions on topics
- Summarize information verbally
- Respond to questions in a timely manner
- Talk about thoughts in an organized way
- Participate in speech similar to an academic discussion

Please indicate your agreement with the following statements. (Strongly agree, Agree, Disagree, Strongly disagree)

- I had enough time to answer the questions on the speaking section.
- The speaking section of the TOEFL iBT let me show how well I can speak in English.
- It is important to include a speaking section on a test of English as a foreign language.

Indicate your level of agreement with the following statements about the integrated tasks on the TOEFL iBT. (Strongly agree, Agree, Disagree, Strongly disagree)

- The integrated tasks on the TOEFL iBT
- Felt natural
- Gave me sufficient time to show my English ability
- Reflected how well I knew the topic before the test
- Let me show my knowledge of culture at English-speaking universities
- Showed me which types of tasks to expect at an English-speaking university
- Showed me what areas of my language ability I should improve before entering an English speaking university

The integrated writing task on the TOEFL iBT that required me to read, listen, and write let me show how well: (Strongly agree, Agree, Disagree, Strongly disagree)

- I can read in English
- I can write in English

- I can listen in English
- I can use academic English vocabulary
- I can use English grammar forms correctly
- I can use English in real-life situations at an English-speaking university

The integrated speaking tasks on the TOEFL iBT that required me to read, listen, and speak let me show how well: (Strongly agree, Agree, Disagree, Strongly disagree)

- I can read in English
- I can speak in English
- I can listen in English
- I can use academic English vocabulary
- I can use English grammar forms correctly
- I can use English in real-life situations at an English-speaking university

In your opinion, is the TOEFL iBT appropriate for the following groups? (Yes/No/No opinion)

- Professionals
- Post-graduates
- College/undergraduates
- Students in all subject areas
- All nationalities/cultures

Now we are going to ask you about the tasks you did at your English-speaking university and how they relate to the questions on the TOEFL iBT.

Indicate your level of agreement with each statement. (Strongly agree, Agree, Disagree, Strongly disagree)

- Preparing for the TOEFL iBT helped prepare me for Using English in an English-speaking university.
- The questions on the TOEFL iBT asked me to use English in ways that I have had to use English at my English-speaking university.

What was your best overall score on the TOEFL Internet-based Test (iBT)?

- 69 or less
- 70 – 79
- 80 – 89
- 90 – 99
- 100 – 109
- 110 or greater
- I don't know or remember.

Appendix D
Instructor Survey

First, we'd like to know more about you and your teaching background.

How many years have you taught English as a Second/Foreign Language (ESL/EFL)

- 1 year or less
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- 7 years or more

Next, we'd like to know about your current position and program.

Which best describes your current position?

- Instructor
- Private tutor
- Both an instructor and private tutor
- Other (please specify)

Where do you currently teach? (Select all that apply.)

- Test preparation center associated with Educational Testing Service (ETS)
- Test preparation center not associated with ETS
- Private ESL/EFL institution
- University-based ESL/EFL institution
- Tutoring company
- Self-employed
- Other (please specify)

How would you characterize your ESL/EFL teaching program? (Select all that apply.)

- A university program for students who are currently enrolled in non-ESL university classes
- A program for students who wish to enter a university
- Other (please specify)

Do you teach course(s) with "TOEFL" in the title?

- Yes
- No

Which of the following best describes your TOEFL classes? (Select all that apply.)

- A constant cycle of TOEFL iBT skills classes
- A TOEFL iBT preparation class held for a set number of weeks
- Students enroll until they receive their desired score on the TOEFL iBT
- Other (please specify)

Are your TOEFL classes designed exclusively for TOEFL preparation?

- Yes
- No, TOEFL preparation is only one component of the class

The purpose of this survey is to find out about your experience preparing students to take the TOEFL iBT. In this section of the survey, we will ask you questions about the test and how you prepare students.

How familiar are you with the TOEFL iBT? Select the best option:

- Very familiar
- Somewhat familiar
- Somewhat unfamiliar
- Very unfamiliar

Do you have adequate access to ETS-produced materials/ information about the TOEFL iBT?

- Yes
- No
- Why or why not? (open-ended response)

Think about students preparing to take the TOEFL iBT. How useful is it for these students to practice the following skills in relation to the TOEFL iBT? (Useful, Somewhat useful, Not very useful, Not useful at all, Don't know)

- Reading
- Writing
- Listening
- Speaking

- Grammar
- Vocabulary
- Practice TOEFL iBT tests
- How to answer multiple-choice questions
- Test-taking skills (outlining main ideas, taking notes, etc.)
- Integrated tasks (tasks that combine language skills: listening, reading, writing, and speaking)
- How to specifically take the TOEFL iBT (i.e., timed speaking, answer formulas for questions, etc.)
- Other (please specify)

What is the most important skill to teach students preparing to take the TOEFL iBT? (open-ended response)

Please rank the sections of the TOEFL iBT based on the importance of preparing for them prior to the test. (Most important, Second most important, Third most important, Least important)

- Listening section
- Reading section
- Writing section
- Speaking section

Do your students report that any of the following factors affect their performance on the TOEFL iBT? (Many, Some, Not many, None)

- Time pressure
- Length of the test
- Students' fear of tests
- Unfamiliarity of topics
- Distraction caused by other test takers
- Difficulty of language on the test
- Computer, mouse, microphone, headphones, etc.

Think about how you prepare students for the listening section of the TOEFL iBT. Approximately how often do you incorporate the following tasks into your instruction? (Often, Sometimes, Rarely, Never)

- Take notes
- Listen for speaker's purpose
- Understand the relationships between ideas
- Listen for introductions, topic changes, and conclusions
- Draw conclusions based on what is implied in the listening
- Make connections among pieces of information in a conversation or lecture

Please indicate your level of agreement with the following statement: (Strongly agree, Agree, Disagree, Strongly disagree)

- The listening section on the TOEFL iBT allows students to show how well they can listen in English.

Think about how you prepare students for the speaking section of the TOEFL iBT. Approximately how often do you incorporate the following tasks into your instruction? (Often, Sometimes, Rarely, Never)

- Pronunciation
- Give oral presentations
- Practice timed speaking
- Hold group discussions/debates
- Make a point and provide supporting examples
- Read about a topic and then talk about it
- Practice formulas for responding to questions on the TOEFL iBT
- Listen to a lecture or conversation and then talk about it
- Find opportunities (such as field trips) for students to speak with native speakers of English

The speaking section of the TOEFL iBT includes integrated speaking tasks which require students to speak in English about a topic they have learned about through a reading and listening task.

Indicate your level of agreement with the following statements. These integrated speaking tasks show how well students can: (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- Read in English
- Speak in English
- Listen in English
- Use appropriate pronunciation
- Use English grammar forms correctly
- Use English in real-life situations at an English-speaking university

Please indicate your level of agreement with the statements about the speaking section below. (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- The speaking section on the TOEFL iBT allows students to show how well they can speak in English.
- The integrated tasks allow students to show how well they can speak in a real-life setting.
- The speaking section of the TOEFL iBT contains real-world tasks.
- Integrated speaking questions show English ability better than questions that measure one skill at a time.
- Students' scores on the speaking section are an accurate reflection of their speaking abilities.
- Students have a sufficient amount of time to formulate responses to the speaking tasks on the TOEFL iBT.
- Students who perform well on the speaking section are prepared for interacting in an English-speaking university.

Think about how you prepare students for the writing section of the TOEFL iBT. Approximately how often do you incorporate the following tasks into your instruction? (Often, Sometimes, Rarely, Never)

- Build vocabulary
- Write opinion essays
- Use idiomatic expressions
- Organize a cohesive essay
- Follow spelling conventions
- Use a range of grammatical structures
- Practice formulas for structuring an essay
- Write about what students have learned from a listening or reading passage

The writing section of the TOEFL iBT includes integrated writing tasks which require students to write about a topic they learned about through a reading and listening task.

Indicate your level of agreement with the following statements. These integrated writing tasks show how well students can: (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- Read in English
- Write in English
- Listen in English
- Use academic English vocabulary
- Use English grammar forms correctly
- Use English in a real-life situation at an English-speaking university

Please indicate your level of agreement with the following statement: (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- The writing section on the TOEFL iBT allows students to show how well they can write in English.
- The integrated tasks allow students to show how well they can write in a real-life setting.
- Integrated writing questions that combine skills show English ability better than questions that measure one skill at a time.

Please indicate your level of agreement with the following statements. (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- I understand what the TOEFL iBT scores mean.
- Users of TOEFL iBT scores look at subscores as well as total scores.
- ETS adequately disseminates information about changes to the TOEFL iBT.
- ETS adequately disseminates information about the meaning of TOEFL iBT scores.
- Preparing to take the TOEFL iBT prepares students for life at an English-speaking university.
- Users of TOEFL iBT scores (admissions officials) understand how to use TOEFL iBT scores.
- The TOEFL iBT is an accurate predictor of how well a non-native English speaker will perform in an English-speaking university.

Is the TOEFL iBT appropriate to students' future English language needs: (Yes/No/Don't know)

- At the pre-university level
- At the undergraduate level
- At the graduate level
- For vocational studies

Please indicate your level of agreement with the statements about the speaking section below. (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- Yes
- It's appropriate for some students
- No
- Don't know

Please use the space below to clarify your response.

Appendix E
Administrator Survey

In which type of institution do you currently work?

- trade or vocational school
- community or junior college
- private college or university (4-year)
- public college or university (4-year)
- Other (please specify)

In which state is your institution located?
Approximately how many international students does your institution serve each academic year?

- 100 international students or fewer
- 101200 international students
- 201 – 500 international students
- 501 – 1,000 international students
- 1,001 international students or more
- Don't know

What is your current job title? (open-ended response)
How long have you worked in your current position?

- 1 year or less
- 2 years
- 3 years
- 4 years
- 5 years
- 6 years
- More than 6 years

How long have you worked with international students as an administrator?

- 1 year or less
- 2 – 3 years
- 4 – 6 years
- 7 years or more

As part of your position, which of the following activities do you perform related to international students? (Select all that apply.)

- Recruit students
- Review applications
- Answer admissions questions
- Advise admitted students
- Place international students
- Build students' English language skills
- Give input on international admissions policies
- Make decisions about international admissions policies
- Other (please specify)

Please answer the following questions about admissions requirements for your institution.

Which of the following tests does your institution accept to meet entrance requirements for English language proficiency? (Yes, No, Don't know)

- International English Language Testing System (IELTS)
- London Tests of English
- Michigan English Language Institute College Entrance Test (MELICET)
- STEP - The Eiken Test in Practical English Proficiency
- Test of English as a Foreign Language paper-based test (TOEFL pBT)
- Test of English as a Foreign Language Internet-based test (TOEFL iBT)
- Test of English for International Communication (TOEIC)
- Test of Spoken English (TSE)
- Other (please specify)

At some institutions, students can meet English language proficiency requirements without submitting an English-language test score.

How does your institution use TOEFL iBT scores? (Select all that apply.)

- Entrance requirement
- Student placement
- Other (please specify)

What is the minimum total score on the TOEFL iBT that is required to enter your institution?

- Under 70
- 70 – 90
- 91 – 105
- 106 or above
- No minimum score required
- Don't know
- Other (please specify)

Please indicate your level of agreement with the following statements. (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- I am familiar with the content of the TOEFL iBT.

- The content of the TOEFL iBT reflects what students need to be able to do at a university in an English-speaking country.
- I understand how the TOEFL Internet-based test is different from the TOEFL paper-based test (pBT).
- I look at both composite and subscores on the TOEFL iBT. I understand how to interpret TOEFL iBT scores.
- The test publisher disseminates adequate information about the meaning of TOEFL iBT scores.
- The TOEFL iBT is a good predictor of how well international students will perform at my institution.
- I am confident using TOEFL iBT scores to make admissions decisions.

Indicate your level of agreement with the following statements. (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- I am familiar with how international students prepare to study in an American university.
- I am familiar with the way international students prepare for the TOEFL iBT.
- Students who have taken the TOEFL iBT have better language skills than those who have not.
- Candidates need to prepare for the TOEFL iBT using materials designed specifically for the TOEFL iBT.
- Candidates have a better chance of getting a good score on the TOEFL iBT if they attend a preparation course.

The TOEFL iBT is a good measure of English language proficiency: (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- At the pre-university level
- At the undergraduate level
- At the graduate level
- For vocational studies

Indicate your level of agreement with the following statements. (Strongly agree, Agree, Disagree, Strongly disagree, Don't know)

- Students from some cultural groups have higher scores on the TOEFL iBT than students from other cultural groups.
- The TOEFL iBT is a fair measure of English language ability for all populations of international students.

Additional Comments (open-ended response)

Appendix F
Selected Stimulated Recall Questions

Interview

Instructions: After each section of the test, you will ask the examinee questions about the section. You will ask both general questions (e.g., "What did you think of this part of the test") as well as a set of questions for each of the questions in the section (e.g., "Do you think this question is a good way to show how well you know English?"). The interview will be audio recorded for you to transcribe, and the protocol also includes space for you to take notes. First, I'm going to tell you about what we'll be doing. You are going to take only a few questions from each of the four sections of the TOEFL Internet-based tests. While you take each section, I'll observe you and take notes. Then, after each section, we'll stop and I'll ask you some questions.

The following questions were asked during the reading section of the Stimulated Recall and were paralleled in the other three modalities.

- Overall, what did you think of this reading section?
- What were you supposed to do in this reading section? (PROBE: What skills did you use in this reading section?)
- Do you read passages like this in school? (If no: What kinds of things do you read in school? How is this different?)
- Do you think the passage is similar to things students have to read when studying at an English-speaking university?
- Take a minute now to look at this question. What were you supposed to do here?
- Do you think this question is a good way to show how well you know English? Why or why not?
- Do you think this question is similar to things you might do in an English-speaking university? Why or why not?

- Now think about all of the questions you answered in the reading section. Overall, do you think the reading section let you show how well you know English? Why or why not?
- Do you think this part of the test let you show how well you would do at an English-speaking university? (If Yes: What kinds of things did you do in this part that were similar to things you would do at an English-speaking university? If No: What kinds of things did you do that aren't similar to what you would do at an English-speaking university?)
- How easy or hard was this part of the test? (PROBE: Can you tell me more? What made it easy/hard for you?)
- Think about the test developers, or the people who created this test. Why do you think they included this reading section on the iBT? (PROBE: Do you think their goals were successful?)

When the participant had completed the Stimulated Recall, they were asked the following global questions:

- Think about all of the sections you completed today. Do you think the iBT shows how well someone could use English at an English-speaking university?
- Do you think it is important for the test to include the types of questions you did today?

Suggested citation:

Malone, M. E., & Montee, M. (2014). *Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability* (TOEFL iBT Report No. 22, ETS Research Report No. RR-14-42). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12039

Action Editor: Gary Ockey

Reviewers: This report was reviewed by the Research Subcommittee of the TOEFL Committee of Examiners.