

*TOEFL*<sup>®</sup> Research **INSIGHT**

# Using *TOEFL iBT*<sup>®</sup> Test Scores for Selecting International Teaching Assistants

VOLUME 10

# **TOEFL® Research Insight Series, Volume 10: Using TOEFL iBT® Test Scores for Selecting International Teaching Assistants**

## **Preface**

The *TOEFL iBT*® test is the world's most widely respected English language assessment and used for admissions purposes in more than 150 countries, including Australia, Canada, New Zealand, the United Kingdom, and the United States (see test review in Alderson, 2009). Since its initial launch in 1964, the *TOEFL*® test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT* test, was launched in 2005. It contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments.

In addition to the *TOEFL iBT* test, the *TOEFL*® Family of Assessments was expanded to provide high-quality, English proficiency assessments for a variety of academic uses and contexts. The *TOEFL*® Young Students Series features the *TOEFL Primary*® and *TOEFL Junior*® tests, which are designed to help teachers and learners of English in school settings. In addition, the *TOEFL ITP*® program offers colleges, universities, and others affordable tests for placement and progress monitoring within English programs as a pathway to eventual degree programs.

At ETS, we understand that scores from the *TOEFL* Family of Assessments are used to help make important decisions about students, and we would like to keep score users and test takers up-to-date about the research results that help assure the quality of these scores. Through the publication of the *TOEFL*® *Research Insight Series*, we wish to communicate to the institutions and English teachers who use any/all of the *TOEFL* tests the strong research and development base that underlies the *TOEFL* Family of Assessments and demonstrate our continued commitment to research.

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

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

internationally recognized leaders in diverse areas such as test validity, language learning and assessment, and educational measurement.

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

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

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

**John Norris, Ph.D.**

Senior Research Director  
English Language Learning and Assessment  
Research & Development Division  
ETS

The following individuals contributed to this volume (in alphabetical order): Ikkyu Choi, Marian Crandall, Ching-Ni Hsieh, John Norris, Spiros Papageorgiou, and Jonathan Schmidgall (lead author).

## Using *TOEFL iBT*® Test Scores for Selecting International Teaching Assistants

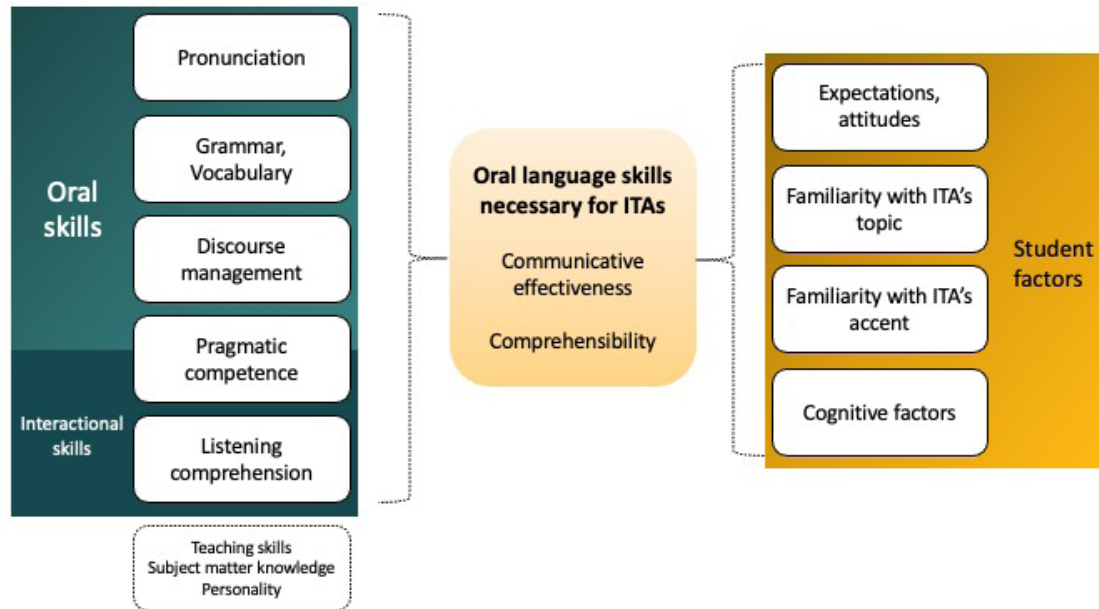
Many universities need to evaluate the oral proficiency of international graduate students prior to their appointment to teaching positions. To do so, universities may develop training and testing programs for International Teaching Assistants (ITAs) but, even when these testing programs exist, practical constraints may make them challenging to administer. Developing a local test that provides reliable, fair, and meaningful information about ITAs' language skills requires expertise in linguistics and educational measurement. Even once a local test is developed, maintaining it may be resource intensive. A coordinator is needed to oversee the ongoing administration of the test, conduct quality assurance activities, and review and communicate certification policies to individual candidates, departments, and school administrators. And a high-quality pool of expert raters needs to be trained and maintained. Given the costs and requirements of running a local testing program at scale, universities may benefit from using *TOEFL iBT* test scores in order to facilitate the screening and selection of ITA candidates.

This volume in the *TOEFL Research Insight Series* proposes approaches that can be taken to establish, monitor, and evaluate score requirements on the *TOEFL iBT* test — often referred to as “cut scores” — to meet institutional mandates and promote ITA success. We begin with a brief summary of the oral language skills needed by ITAs, as indicated by previous research and local ITA testing programs. We then discuss the language skills measured by the *TOEFL iBT* test and how they relate to the language skills needed by ITAs. We conclude by reviewing the ways in which *TOEFL iBT* test scores may be used to support ITA screening decisions. This review includes practical recommendations for setting cut scores and suggestions for evaluating the efficacy of cut scores.

### The oral language skills needed by ITAs

Although ITAs use all four language skills (listening, reading, speaking, writing) as part of their duties, language certification policies for ITAs have typically emphasized oral language skills (Farnsworth, 2014). Figure 1 shows how the ultimate goal of ITA certification — ensuring that an ITA has the oral skills necessary to work as a Teaching Assistant (TA) — is influenced by an ITA's oral skills and factors related to undergraduate students (see Schmidgall, 2013).

**Figure 1 Summary of the predictors of effective communication for ITAs**



The left side of Figure 1 shows that the language skills needed by ITAs to communicate effectively (or comprehensibly) include oral skills such as pronunciation, knowledge, and effective use of grammar and vocabulary; the ability to structure and manage discourse; and pragmatic competencies. In addition, ITAs need to use listening skills in interaction. Research has also shown that other attributes of ITAs — such as teaching skills, subject matter knowledge, and personality — may impact judgments of how effectively an ITA can communicate. The right side of Figure 1 shows that evaluations of an ITA’s communicative effectiveness are influenced by the characteristics of those who listen or interact with them; typically, undergraduate students. Students may vary in terms of their expectations and attitudes toward ITAs, their familiarity with the content or topic of discussion, their familiarity with different accents, and their cognitive factors such as working memory capacity or fatigue. Although ITA certification focuses on evaluating an ITA’s skills, it is important to be aware that the extent to which ITAs are perceived as effective communicators may also be influenced by characteristics of the students.

Another way to examine the language skills needed by ITAs is by reviewing the skills typically evaluated by ITA assessments developed internally by universities. Although these assessments are developed, administered, and scored with practical constraints in mind (e.g., time, money, expertise), they reflect how ITA proficiency is conceived and ITA policy is implemented by institutions. Schmidgall (2013) found that ITA assessments in North American universities typically measure oral skills (pronunciation, fluency, vocabulary, grammar, rhetorical organization, comprehensibility) and interactional skills (question handling or listening and responding appropriately in interaction). These ITA assessments rarely measure non-linguistic skills — such as teaching effectiveness — because of the consideration that even native speakers of English are not required to demonstrate teaching skills prior to TA appointments (Farnsworth, 2014). ETS researchers conducted a review

of 62 North American universities' ITA assessment programs in Fall 2019 and found that 37 of these programs (60%) used some form of local assessment, often in conjunction with large-scale assessments. Although local ITA assessments vary in terms of the types of tasks and the components of language proficiency that are explicitly evaluated, the emphasis on oral and interactional skills largely aligns with the research-based summary shown in Figure 1. The tasks included in ITA assessments reflect the tasks ITAs are required to perform, including giving clear lectures and responding to student questions.

### **The relevance of the TOEFL iBT test in the context of ITA screening**

In the review of the 62 local ITA assessment programs mentioned earlier, the researchers found two major ways in which policies varied:

- the range and types of certification decisions, and
- how local ITA assessments and scores from the TOEFL iBT test contributed to certification decisions.

Some ITA programs only included two types of certification decisions — Pass and Fail — while others included up to three conditional Pass decisions that authorized ITAs to be placed into specific types of TA positions, possibly pending additional language support. In addition to this variation in terms of the types of certification decisions, ITA programs varied in how they incorporated their own assessments versus the TOEFL iBT test into their certification policies.

Since international graduate students may have taken the TOEFL iBT test as part of the admissions process, ITA program administrators may want to incorporate this pre-existing information about language proficiency into ITA certification procedures. Given the potential usefulness of incorporating TOEFL iBT test scores into university ITA policies, there are two critical issues to consider. The first issue is whether TOEFL iBT test scores provide relevant information about ITA language proficiency in the local context. The second issue is the extent to which TOEFL iBT test scores predict desirable outcomes — such as certification decisions by local ITA tests, or language or teaching evaluations of ITAs.

Related to these two issues, it is important first to understand what the TOEFL iBT test is intended to measure and then to illustrate the relationship of what is measured on the TOEFL iBT test to ITA language use. As described in *Volume 1: TOEFL iBT® Test Framework and Test Development*, the TOEFL iBT test was designed to measure examinees' English language proficiency in situations and tasks that are reflective of university life where instruction is conducted in English. Consequently, the design of the TOEFL iBT test incorporates a purposeful consideration of competencies and tasks relevant to academic language use. Figure 2 summarizes these competencies and tasks for the speaking and listening sections of the TOEFL iBT test because speaking and listening skills are necessary for communication in the classroom.

**Figure 2 The constructs measured by the TOEFL iBT Speaking and Listening subtests**

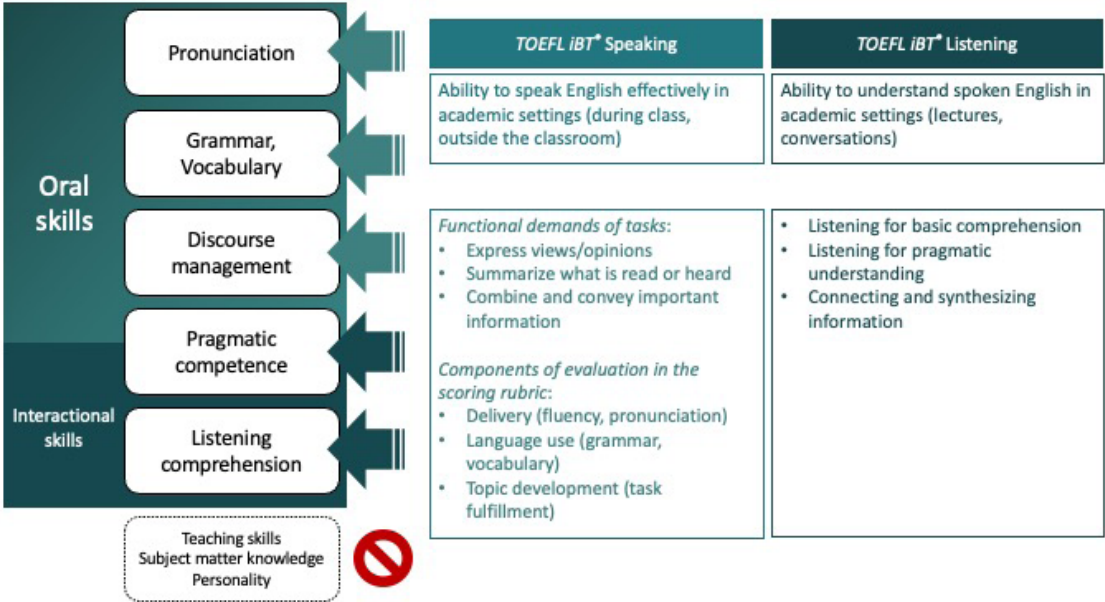
	TOEFL iBT® Speaking	TOEFL iBT® Listening
<b>Purpose</b>	To evaluate the ability to speak English effectively in academic settings (classroom, campus encounters)	To evaluate the ability to understand spoken English in academic settings (classroom, campus encounters)
<b>Tasks</b>	<ul style="list-style-type: none"> <li>• Independent</li> <li>• Integrated</li> </ul>	<ul style="list-style-type: none"> <li>• Listen to conversations</li> <li>• Listen to lectures</li> <li>• Respond to comprehension questions</li> </ul>
<b>Language knowledge and skills (Competencies)</b>	<p><i>Functional demands of tasks:</i></p> <ul style="list-style-type: none"> <li>• Express views/opinions</li> <li>• Summarize what is read or heard</li> <li>• Combine and convey important information</li> </ul> <p><i>Components of evaluation in the scoring rubric:</i></p> <ul style="list-style-type: none"> <li>• Delivery (fluency, pronunciation)</li> <li>• Language use (grammar, vocabulary)</li> <li>• Topic development (task fulfillment)</li> </ul>	<ul style="list-style-type: none"> <li>• Listening for basic comprehension</li> <li>• Listening for pragmatic understanding</li> <li>• Connecting and synthesizing information</li> </ul>

As shown in Figure 2, the speaking section of the TOEFL iBT test was designed to evaluate a candidate’s ability to speak English effectively in academic settings. The test evaluates speaking proficiency using a task-centered design that includes independent and integrated speaking tasks. Each task has slightly different functional demands associated with it, and these functional demands involve expressing views or opinions, summarizing what is read or heard, or combining and conveying important information. Candidates’ performances on these tasks are evaluated based on their delivery, language use, and topic development. The listening section of the TOEFL iBT test was designed to evaluate a candidate’s ability to understand spoken English in academic settings. Through the task-and-competency-based test design, candidates listen to lectures or conversations and answer comprehension questions. In order to complete these tasks, candidates draw upon their competencies related to listening comprehension. This includes listening for basic comprehension or understanding the main idea, major points, and important details related to the main idea. Another competency is listening for pragmatic understanding, which involves recognizing a speaker’s attitude and degree of certainty, as well as the function or purpose of what a speaker says. In addition, tasks may require connecting and synthesizing information, which involves:

- recognizing the organization of information presented,
- understanding the relationship between ideas presented,
- making inferences and drawing conclusions based on what is implied,
- making connections among pieces of information in a conversation or lecture, and
- recognizing topic changes in lectures and conversations, and recognizing introductions and conclusions in lectures.

As shown in Figure 3, there is obvious overlap between the language skills needed by ITAs (Figure 1) and the skills evaluated in the speaking and listening sections of the TOEFL iBT test (Figure 2). The speaking section incorporates an evaluation of pronunciation, grammar and vocabulary, and discourse management (as indicated by the blue arrows). The listening section evaluates listening comprehension, understanding of grammar, vocabulary, and discourse (as indicated by the red arrow). The listening section also evaluates, albeit to a more limited extent, pragmatic competence through the evaluation of pragmatic understanding (as indicated by the red and white arrow). As with many locally-developed ITA assessments, the TOEFL iBT test is not intended to measure non-linguistic abilities such as teaching skills, subject matter knowledge, and personality-related factors. In addition, although TOEFL iBT test tasks are set in academic settings, they are not oriented towards the instructor’s role, which can potentially influence how these skills are used or realized in an instructional context.

**Figure 3 Relationship between the language skills measured by the TOEFL iBT test and the language skills needed by ITAs**



Relevant research shows that the design of the TOEFL iBT test engages oral skills that are highly relevant to TA duties. Cotos and Chung (2018) found that test-taker responses to TOEFL iBT test speaking tasks contained language functions similar to those used by ITAs. For example, Cotos and Chung analyzed a corpora of real-world ITA discourse and found that description (e.g., describing, comparing, exemplifying), principles (e.g., explaining, predicting, hypothesizing), and sequence (e.g., reporting, instructing, narrating) structures were relatively common. These structures were also relatively common in test takers’ responses to TOEFL iBT test speaking tasks. Farnsworth (2013) concluded that the speaking section of the TOEFL iBT test and a university’s local ITA test measured the same underlying language ability. Xi’s (2007, 2008) research showed that speaking scores from the TOEFL iBT test could be used effectively to screen high-level and low-level ITA candidates.



The ultimate goal of ITA certification is to help ensure that ITAs are able to communicate effectively with undergraduate students. Consequently, student perceptions or judgments of an ITA's comprehensibility or communicative effectiveness are an important, real-world outcome measure to consider when using the TOEFL iBT test to evaluate the language proficiency of ITA candidates. Bridgeman, Powers, Stone, and Mollan's (2011) study showed that the speaking scores of the TOEFL iBT test are strongly correlated with undergraduates' judgments of test takers' comprehensibility. In another study, Wagner (2016) investigated the extent to which TOEFL iBT speaking and listening test scores predicted student evaluations of ITAs. The results of this study suggested that the ITA's listening ability also plays a part in student evaluations and that listening scores from the TOEFL iBT test can be a useful measure to incorporate into ITA screening.

To briefly summarize, research provides evidence that TOEFL iBT test scores can be a useful indicator of the language skills necessary for ITAs and that the speaking and listening sections, in particular, may be useful for supporting ITA screening decisions.

### **Establishing, monitoring, and evaluating TOEFL iBT test cut scores for ITA screening and placement decisions**

English language proficiency is one of several important factors to consider when determining whether international graduate students are prepared to accomplish the various tasks associated with teaching at U.S. universities. The TOEFL iBT test provides a trustworthy indication of a test taker's English language proficiency in each of the four skills of listening, speaking, reading, and writing. It was developed and validated specifically in relation to language use that is typical of university academic settings. As the previous sections have shown, the TOEFL iBT test can serve as an appropriate tool for screening a teaching candidate's English language abilities and many U.S. universities use the test in this way.

This section provides suggestions for using TOEFL iBT scores in the ITA screening process, including:

- a sample of U.S. universities and their practices,
- conducting a standard setting study to determine appropriate TOEFL iBT cut scores,
- an initial recommendation for cut scores for each section of the TOEFL iBT test, and
- approaches to monitoring and evaluating the effectiveness and accuracy of the TOEFL iBT test scores used for ITA screening.

### **A sample of U.S. universities and their practices for using TOEFL iBT scores for ITA screening**

In Fall 2019, ETS researchers conducted an internet search to identify the current policies used for ITA screening and assessment at U.S. universities. In total, the researchers reviewed the websites at 62 universities — which varied greatly in terms of the amount of information they provided about their ITA screening policies and assessments — and they were able to identify 40 ITA screening policies that utilized TOEFL iBT test scores.

In general, these screening policies utilized TOEFL iBT test scores in two ways. The most common use was to screen ITA candidates at the highest level, typically characterized as "Pass" or "Full Pass." The median TOEFL iBT Speaking test cut score used for "Pass" decisions was 26 and ranged from 22 to 28 across the university policies surveyed.

A smaller number of universities' policies (9 of 40) also used TOEFL iBT test scores to screen at additional levels. Policies greatly varied in the extent to which they included "Conditional Pass" levels, as well as how these levels and policy implications were defined. For example, one ITA policy included three Conditional Pass levels based on an ITA's level of oral proficiency and those levels had different implications for the kinds of instructional activities authorized for those candidates as well as the additional language training requirements imposed. Conditional pass levels might be useful for programs that need more flexibility in their ITA testing program, as they allow ITA certification for narrower purposes or with practical remedial requirements.

### **Conducting a study to determine appropriate TOEFL iBT test cut scores for ITA screening**

The most effective way to establish requirements for ITAs using TOEFL iBT test scores is to conduct a standard setting study. Standard setting uses a combination of expert judgment and empirical methods to identify the minimum level of proficiency needed to support a policy decision. There are a variety of methods that can be used to conduct such a study; one example is provided by Wylie and Tannenbaum (2006). The purpose of this study was to establish cut scores for screening ITAs using scores on the speaking section of the TOEFL iBT test. A panel of 18 experts who work with ITAs was convened to judge the meaning and relevance of two different cut scores. First, the researchers determined a cut score for minimally acceptable speaking skills for ITAs to be able to work with undergraduate students, which they set at a score of 23 out of 30 points on the speaking section of the TOEFL iBT test. Second, the researchers determined a higher cut score based on the statistical relationship between the speaking section of the TOEFL iBT test and the *Test of Spoken English (TSE)*; this higher level was intended to indicate speaking proficiency that is fully adequate for teaching purposes. The researchers set this higher-level cut score at 26 out of 30 points on the speaking section of the TOEFL iBT test. The findings of this study corroborate the findings of other studies and the review of many current ITA policies presented above, and they provide a useful basis for initial recommendations regarding appropriate cut scores for screening ITAs (see next section).

The literature also points out that the same cut score might not be equally relevant across different ITA contexts. When deciding to allow an international student to teach, institutions need to consider a variety of factors, such as:

- The type of classification error that is deemed more important to avoid, such as a false positive (allowing a student to teach when the student does not have sufficient language proficiency) or a false negative (not allowing students with sufficient language proficiency to teach). Both classification errors have important consequences. False positive classifications are likely to result in frustration among undergraduate students taught by the prospective ITA, whereas false negative errors are likely to have consequences at the institutional level because departments might not be allowed to employ qualified graduate students.
- The remedial measures that an institution has in place. For example, students who receive a score below the cut score might be asked to take an English language course and teach upon successful completion of that course; in this case some false positive classifications can be rectified. If, however, the institution does not have enough resources to support such a language course, then a higher cut score might be required to minimize false positive classifications.

- The importance of distinguishing between sufficient language proficiency and adequate teaching skills. A language test such as the TOEFL iBT test should only be used as an indicator of the former, not the latter. In cases where undergraduate students complain about an ITA, for example, institutions need to examine whether such complaints are due to a lack of sufficient English language skills versus teaching skills and intercultural competence.

### A baseline recommendation for using TOEFL iBT scores for ITA screening

Cut scores on the TOEFL iBT test for screening ITAs are likely to vary across institutions because of various contextual characteristics, such as the language demands of the instruction, amount of language support offered, and other reasons. Table 1 provides baseline recommendations based on current practice and score-related data.

**Table 1 Baseline recommendations for TOEFL iBT test score requirements for screening ITAs**

Test section	Recommended cut scores (on a scale of 0–30)
Speaking	<ul style="list-style-type: none"> <li>• 27–30 (unconditional appointment)</li> <li>• 23–26 (conditional appointment)</li> </ul>
Listening	23
Reading	24
Writing	23

Based on research findings and current practices of universities reviewed in this volume, speaking cut scores can be recommended for the following decisions:

- Unconditional appointment: 27 to 30 (No additional language testing or training is required before teaching.)
- Conditional appointment: 23 to 26 (Additional requirements will be determined by the institution, for example:
  - ◆ a student delivers a teaching demonstration to a panel for determination regarding readiness to use English in the classroom,
  - ◆ a student begins teaching but is monitored closely during the initial weeks of teaching, or
  - ◆ a student participates in additional language instruction, such as a course on English for teaching purposes.)
- Non-appointment: 22 or below (Teaching in English is not permitted until speaking proficiency is improved. The criterion for improvement is set by the institution.)

Although ITA certification policies typically focus on oral language skills (as discussed earlier), there are several reasons why policymakers may want to consider incorporating information about literacy (reading and writing) skills. First, all four English language skills may be important for teaching and teaching-related work typically done by ITAs given the prevalence of e-mail communication and the role of ITAs in grading and providing written feedback. Second, research has shown that an evaluation of a specific language skill (e.g., speaking) can be improved by assessing related skills (Powers & Powers, 2015).

With these considerations in mind, a minimum threshold of proficiency for reading, listening, and writing may be identified in reference to typical performance expectations for ITAs. As discussed in Volume 9: *Guidelines for Setting Useful Score Requirements for the TOEFL iBT Test*, the minimum listening and reading scores for the advanced level are 22 and 24 respectively<sup>1</sup>. The percentile rank for these scores in 2019 (48 and 57 respectively<sup>2</sup>) indicates that an ITA candidate would have scored around the 50th percentile of all graduate students taking the TOEFL iBT test; in other words, about half of TOEFL iBT test takers would have met the score requirements for ITAs if these scores were used for listening and reading. For writing, a score of 22 indicates that the examinee has scored around the 50th percentile of all graduate students, while a score of 23 suggests an incremental improvement to around the 60th percentile. This score is also the top score for the higher intermediate level.

### **Approaches to monitoring and evaluating cut scores and decisions**

Any use of test scores for making important decisions, including screening of ITA candidates, should be monitored and evaluated for accuracy, effectiveness, and consequences. Many possibilities exist for local validation of test use along these lines, several of which are presented here as examples of suggested good practice.

- Analyze the proportion of ITAs screened into the different categories proposed in the section on baseline recommendations for using TOEFL iBT scores in comparison with the ITA needs of departments and their existing placement practices. Consider if the use of TOEFL iBT test cut scores are similar, more severe, or more lenient. And think about the implications for departments, students, and English language support based on initial screening decisions.
- Survey and/or interview ITAs placed into teaching positions at the midpoint and end of the semester in order to determine their perceptions regarding the English language demands of teaching and their readiness in terms of each of the four skills.
- Survey students in ITA-taught classes at the end of semester in order to determine their perceptions regarding the English language demands of teaching and their teachers' readiness in terms of general-language proficiency and/or each of the four skills (possibly in combination with regular end-of-semester course evaluations). When surveying students, it is important to keep in mind some of the individual factors that may influence their perceptions of their teacher's proficiency, as shown previously in Figure 1.

---

<sup>1</sup> See all level descriptions for TOEFL iBT test scores at [www.ets.org/toefl/institutions/scores/interpret](http://www.ets.org/toefl/institutions/scores/interpret).

<sup>2</sup> The percentile rank of a score indicates the percentage of test takers at that score and below. The TOEFL program provides information about percentile ranks for all tests administered in a calendar year in the annual *TOEFL iBT® Test and Score Data Summary* ([www.ets.org/toefl/score-users/resources-services](http://www.ets.org/toefl/score-users/resources-services)).

- Survey departmental decision makers regarding graduate students placed into teaching positions, as well as non-appointed graduate students in order to determine their perceptions of the number of:
  - ◆ accurate screening decisions,
  - ◆ false positives, and
  - ◆ false negatives.
- Conduct regular classroom (and non-classroom instructional work, as applicable) observations of newly placed ITAs to assess “on-the-job” use of English skills.
- Compile evidence and reflect on initial cut scores after one or several iterations of use; consider possible adjustments in cut scores and the likely consequences.

## References

- Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, 26(4), 621–631. doi:10.1177/0265532209346371
- Cotos, E., & Chung, Y.-R. (2018). Domain description: Validating the interpretation of the TOEFL iBT speaking scores for international teaching assistant screening and certification purposes (TOEFL Research Report No. RR-85). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12233>
- Educational Testing Service. (2018). *Validity evidence supporting the interpretation and use of TOEFL iBT scores* (Volume 4). Retrieved from [https://www.ets.org/s/toefl/pdf/toefl\\_ibt\\_insight\\_s1v4.pdf](https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf)
- Farnsworth, T. (2013). An investigation into the validity of the TOEFL iBT Speaking test for International Teaching Assistant certification. *Language Assessment Quarterly*, 10, 274–291.
- Farnsworth, T. F. (2014). Assessing the oral abilities of international teaching assistants in the USA. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 471–483). New York, NY: John Wiley.
- Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing*, 32(2), 151–167.
- Schmidgall, J. E. (2013). Modeling speaker proficiency, comprehensibility, and perceived competence in a language use domain (Doctoral dissertation, University of California at Los Angeles). Retrieved from <https://cloudfront.escholarship.org/dist/prd/content/qt58g6k6zg/qt58g6k6zg.pdf>
- Wagner, E. (2016). *A study of the use of TOEFL iBT Speaking and Listening scores for international teaching assistant screening* (TOEFL iBT® Research Report No. 27). Princeton, NJ: Educational Testing Service.
- Wylie, E., & Tannenbaum, R. (2006). *TOEFL academic speaking test: Setting a cut score for international teaching assistants* (Research Memorandum No. RM-06-01). Princeton, NJ: Educational Testing Service.
- Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, 4(4), 318–351.
- Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs* (TOEFL iBT® Research Report No. 3). Princeton, NJ: Educational Testing Service.