ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–27
ETS Research Report No. RR–16-18

A Study of the Use of the *TOEFL iBT*® Test Speaking and Listening Scores for International Teaching Assistant Screening

Elvis Wagner

May 2016

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 (2015 – 2016) members of the TOEFL COE are:

Lia Plakans - Chair	University of Iowa
Sara Weigle	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
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

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

RESEARCH REPORT

A Study of the Use of the *TOEFL iBT*® Test Speaking and Listening Scores for International Teaching Assistant Screening

Elvis Wagner

Temple University, Philadelphia, PA

Although the speaking section of the *TOEFL iBT*® test is used by many universities to determine if international teaching assistants (ITAs) have the oral proficiency necessary to be classroom instructors, relatively few studies have investigated the validity of using TOEFL iBT scores for ITA screening. The primary purpose of this study was to determine the effectiveness of using TOEFL iBT Speaking and Listening scores for ITAs in the context of an urban research university. This was accomplished by correlating the TOEFL iBT test scores with the scores obtained by prospective ITAs on local screening tests and student evaluations of ITAs' language and teaching competence and by using the TOEFL iBT test scores as predictors of ITA teaching performance as measured by student evaluations. Given that listening comprehension is an important aspect of instructional language competence, both TOEFL iBT Speaking and TOEFL iBT Listening scores were used as predictors. The development of ITAs' oral proficiency over a semester in an English-speaking environment was also examined by comparing their TOEFL iBT Speaking and Listening scores at the beginning and at the end of their first semester as ITAs. The results indicate that TOEFL iBT Speaking and Listening scores correlate moderately with the local screening exam, but more importantly, the results indicate that TOEFL iBT Listening might be a better predictor of teaching competence. Although TOEFL iBT Speaking scores did not correlate with or predict the various measures of teaching competence, TOEFL iBT Listening scores did. In addition, the study found that ITAs living and teaching in an English-speaking environment did make measurable gains in oral proficiency over the course of a 3-month period. Although these gains were relatively small, and the listening score gains were not statistically significant, this is the first study that has actually tried to measure the improvement in ITAs' oral proficiency while living and studying (and sometimes teaching) on campus. The importance of this study goes beyond the local context. This is the first study to examine the relationship between TOEFL iBT Speaking and Listening scores and instructional performance of ITAs, thus serving as a validation study of the use of the TOEFL iBT test for ITA placement purposes. The study also offers insights into ITAs' development of oral proficiency once they are in an English-speaking environment, which can serve as a basis for future ITA curriculum development.

Keywords International teaching assistants; assessing listening; assessing speaking; test validation; oral language proficiency; teaching competence

doi:10.1002/ets2.12104

To be a competent university instructor in an English-medium university, a person needs to have a certain level of oral proficiency in English. The best way to determine if a nonnative speaker of English has that necessary oral proficiency is an issue with which politicians, university administrators, and English as a second language (ESL) professionals have grappled for decades. Recent research (e.g., Butler, Eignor, Jones, McNamara, & Suomi, 2000; Farnsworth, 2013; Rosenfeld, Leung, & Oltman, 2001; Wylie & Tannenbaum, 2006) has examined the use of the speaking scores of the *TOEFL iBT*® test as a measure of speaking ability in an academic target language use (TLU) domain for international teaching assistants (ITAs). More specifically, Xi (2007, 2008) conducted a criterion-related validity study of the use of TOEFL iBT Speaking scores and locally administered screening tests as measures of ITA placement. Xi established that the relationships between the TOEFL iBT Speaking scores and local ITA screening scores were moderately strong, although the strengths of the relationships differed across the four settings.

The present study sought to investigate the use of the TOEFL iBT scores as an initial screening measure for ITAs at Temple University, while extending Xi's (2007, 2008) research in three ways. First, given that listening comprehension is an important aspect of instructional language competence, this study examined the use of both TOEFL iBT Listening scores

Corresponding author: E. Wagner, E-mail: elviswag@temple.edu

and TOEFL iBT Speaking scores for ITA screening purposes. Second, this study included measures of ITAs' instructional performance (rather than just their performance on local screening tests) to investigate the predictive validity of the TOEFL iBT Speaking and Listening scores. Third, because it is widely assumed that living, working, and studying in an English-speaking environment should lead to significant gains in learners' oral proficiency, assessments of ITAs' oral proficiency (as measured by TOEFL iBT Speaking and Listening scores) at the beginning and at the end of their first semester as ITAs were conducted.

Review of the Literature

Oral Proficiency Requirements of International Teaching Assistants

As a result of complaints from undergraduate students about the lack of oral proficiency of ITAs, a number of states passed legislation in the 1980s and 1990s mandating that all classroom instructors in higher education who are not native speakers of English be certified as having a certain level of English proficiency before they can assume teaching responsibilities (Oppenheim, 1998; Thomas & Monoson, 1991). Numerous studies (e.g., Dick & Robinson, 1994; Monoson & Thomas, 1993; Xi, 2007, 2008) have investigated how different university ITA programs responded to these mandates and the setting of minimal levels of oral proficiency. These studies have found variation in how the different programs respond to the mandates, but the most common model seems to include an initial screening that requires potential ITAs to demonstrate their oral proficiency by scoring at a certain level on a standardized English proficiency test such as the TOEFL iBT test, Pearson Test of English Academic, or International English Language Testing System (IELTS; Farnsworth & Wagner, 2013) as part of the application for admission. Farnsworth and Wagner found that while all four components of the TOEFL iBT were used for admissions purposes, usually only the TOEFL iBT Speaking scores were used for ITA screening purposes. In other words, when the prospective ITA applied to the university as a student, all four of the TOEFL iBT subscores were considered, but only the speaking score was used when that same prospective ITA was screened to be an instructor.

Most second language acquisition researchers and language testers (e.g., Douglas, 2000; Wagner, 2010a, 2014a) would agree that the oral skills of listening and speaking are highly interrelated. Similarly, most educators would agree that teaching involves more than just lecturing to students; rather, effective teaching demands that teachers interact with their students and be able to comprehend and respond to their students' questions and comments (Elder, 1993). Studies that have investigated the teaching that ITAs are required to do invariably include listening and responding to students' questions and engaging in interactive conversations with students as part of the TLU domain (e.g., Gorsuch, 2003, 2006; Hoekje & Linnell, 1994; Papajohn, 2006; Saif, 2002). Myers (1994) described how important it is for ITAs to be able to respond appropriately to students' questions, and Plakans (1997) described how undergraduate students reported being most frustrated by their ITAs' inability to answer the students' questions. Plough, Briggs, and Van Bonn (2010) performed a study investigating the Graduate Student Instructor Oral English Test, which is a language proficiency examination with a teaching TLU domain and is used for ITA screening. They investigated which language abilities predicted success on the test and found that listening comprehension was the greatest predictor of success. Similarly, Elder (1993) investigated how well the IELTS test performed as a predictor of performance in a teacher education program. She found that the IELTS Listening section correlated more highly ($r = .40$) with first-semester progress ratings in the teacher education program than the other IELTS scores (global, reading, writing, speaking). Elder also found that the IELTS Listening scores correlated more highly ($r = .29$) than any of the other IELTS scores with second-semester progress ratings. She reasoned that listening ability is necessary for "survival in the interactive classroom environment" (p. 77). Xi's (2008) study investigated this issue in an ITA context when she analyzed the locally developed teaching simulation tests that were used for ITA selection at four different ITA programs, and all four of these tests included listening comprehension or question handling (or both) as part of the scoring rubrics.

Except perhaps in a very small number of "pure" lecture courses, a competent teacher needs to be able to listen and respond to students' questions and concerns, and listening ability is recognized as an integral component of teaching competence. Yet most ITA programs that rely on large-scale standardized tests (such as the TOEFL iBT or IELTS) to assess potential ITAs' language proficiency seem to have focused exclusively on candidates' speaking scores when judging a prospective ITA's oral proficiency and readiness to teach. In other words, these programs require that ITAs score at a certain level on the speaking sections of these tests but no corresponding minimum listening score are mandated (Farnsworth

& Wagner, 2013). Interestingly, Douglas (1997), in his monograph devoted to issues related to testing speaking ability in academic contexts, questions the usefulness of separating speaking and listening scores, suggesting that the reported scores for the two skills be combined. It should be noted that the speaking section of the TOEFL iBT (described later) does indeed involve both listening and speaking; two of the six speaking tasks require test takers to listen to a text and then provide an oral response. Thus, although separate speaking and listening scores are reported with the TOEFL iBT, at least to some extent, the scores are interdependent. Yet those ITA programs that do use the TOEFL iBT scores for ITA screening purposes seem to use only the speaking scores (Farnsworth & Wagner, 2013). As Powers and Powers (2015) argued, while in some contexts it might be advisable to assess English proficiency in a single area (e.g., speaking), using a more comprehensive assessment approach in which additional skills (speaking, listening, reading, and writing) are assessed can result in better predictions of a test taker's competence in a particular domain. In the ITA teaching domain, because teaching obviously involves both speaking and listening ability, it would seem advantageous to include both speaking and listening TOEFL iBT scores as predictors of teaching competence.

International Teaching Assistant Screening and Preparation

In addition or as an alternative to the TOEFL iBT, many ITA programs require that prospective ITAs take and pass locally administered screening tests as a way of demonstrating oral proficiency in English. As noted earlier, a typical ITA screening process involves all prospective ITAs taking a proficiency exam (e.g., the TOEFL iBT) as part of the application for admission. Generally, the cut score for admission to the university is lower than the cut score for being an ITA. As a result, some potential ITAs are accepted to the university but might not meet the ITA oral proficiency requirement. Xi (2008) described how the rationale of many universities and ITA programs is to set these initial ITA cut scores relatively high to minimize Type 1 errors. The ITA programs want to err on the side of caution in determining the level of English proficiency needed to perform instructional duties; it is better to set the cut scores too high (a Type 2 error) than to set them too low (a Type 1 error). A Type 2 error is harmful to the individual ITA, but a Type 1 error might be harmful to a large number of undergraduate students who would be taught by an ITA whose English proficiency is inadequate. In addition, the Type 2 errors can be mitigated, because even if the ITAs do not meet the high cut scores on these initial screening measures, they will still have a chance to demonstrate their oral proficiency through local screening exams. If the potential ITA does not score high enough on the TOEFL iBT Speaking section, many universities will require the potential ITA to take a local screening exam once he or she arrives on campus. Often, but not always, these local screening exams involve a teaching simulation. If he or she does not meet the cut score on this local screening exam, the potential ITA will usually be required to take and pass an ITA training or preparation course.

The requirements of this ITA training course obviously vary from university to university; often this course is a for-credit course (usually three credits) with a curriculum that focuses on teaching English oral skills as well as good pedagogical practices. Typically there is some sort of culminating assessment that requires the ITA to perform a teaching simulation in a language for specific purposes (LSP) model (Farnsworth, 2013; Farnsworth & Wagner, 2013); that is, the culminating assessments usually require the potential ITAs to demonstrate their oral proficiency through a performance assessment where they teach a simulated lesson in their field to a group of students (Farnsworth, 2013; Schmidgall, 2013). Some universities utilize a stronger LSP test (Douglas, 2000), where teaching competence is part of the evaluation of the assessment, whereas other universities use a weaker version, where the performance test involves a teaching context but teaching competence is not part of the construct being assessed and not part of the scoring rubric. Although most ITA training courses require this culminating teaching simulation assessment, there is a lively debate in the field about whether teaching competence should be part of the assessment (Bailey, 1985; Farnsworth, 2013; Hoekje & Williams, 1992; Saif, 2002; Schmidgall, 2013). In other words, although teaching pedagogical skills seems to be on the curricula of almost all ITA preparation classes, the culminating assessment of these classes may or may not include the assessment of teaching competence. The argument against the inclusion of teaching competence is based on the notion that assessing teaching competence is inherently unfair to nonnative speakers of English, in that native-speaking teaching assistants do not have to demonstrate teaching competence before assuming instructional duties (Bailey, 1985; Farnsworth, 2013; Hoekje & Williams, 1992; Saif, 2002). In contrast, the argument for including teaching competence in the scoring of the culminating assessment seems to be that the ITAs (and their students) would benefit from learning good pedagogical practices, and if pedagogy is part of the curricula for these ITA preparation classes, it is logical to include teaching competence on the final assessment.

One study that investigated the effectiveness of ITA training programs was conducted by Oppenheim (1998). She had 89 prospective ITAs take a teaching performance assessment in which they taught a simulated introductory class in their discipline to a group of undergraduate students. This teaching performance was videotaped. The prospective ITAs then completed a teaching seminar for ITAs that included 24 hours of lectures, workshops, and role plays that focused on teaching goals that fostered active student learning. After completing this preparatory teaching seminar, the ITAs then participated in a second teaching performance assessment in which they were again videotaped teaching a simulated introductory class in their discipline to a group of undergraduate students. A different group of undergraduate students then watched the videotapes of the lessons and scored the ITAs' teaching performance using a teaching effectiveness scale that had four subscales: pedagogy, interpersonal skills, linguistic characteristics, and the students' familiarity with the concepts covered. Paired sample *F*-tests were used to compare scores on the preteaching seminar performance assessments with scores on the postteaching seminar performance assessments. Oppenheim found a relatively small (but statistically significant) improvement in scores on the teaching effectiveness scale. The 89 ITAs scored higher on the assessment after taking the ITA teaching seminar ($M = 3.50$, $SD = .78$) than they did before taking the teaching seminar ($M = 3.13$, $SD = .86$). She concluded that the preparatory ITA teaching seminar did lead to increased teaching ability for the ITAs who took the class.

Halleck and Moder (1995) investigated how the *TOEFL*® test correlates with and predicts performance on a local ITA performance assessment. The ITA performance assessment in their study was the "ITA test," which is very similar to the TEACH test used in the current study. This ITA test involved a teaching simulation, which was then scored on presentation language skills, teaching skills, and interactive language skills. Halleck and Moder found that although the TOEFL test correlated moderately with the three components of the ITA test ($r = .56 – .58$, $p < .05$) and moderately for the total ITA test scores ($r = .57$, $p < .05$), the TOEFL test was not a good predictor of ITA test scores when it was used as a predictor variable (along with oral proficiency interview [OPI] scores) in a linear regression. Whereas the OPI scores accounted for 77% of the variation in total ITA test scores, the TOEFL test only contributed 2%. However, this study was from 1995, when there was not a speaking component to the TOEFL test, and thus it is not particularly surprising that the TOEFL test would not be a good predictor of teaching competence, especially for a teaching competence test that explicitly includes an interactive language skills component.

Two recent and more relevant studies have investigated how the TOEFL iBT correlates with and predicts performance on these local ITA placement performance assessments. Xi (2008) explored the validity of using TOEFL iBT Speaking scores as a screening measure for initial ITA placement by examining the relationship between TOEFL iBT Speaking scores and locally administered ITA placement exams at four different U.S. universities. These locally administered exams varied in their focus. Some attempted to measure test takers' oral proficiency only, whereas others were broader assessments of teaching preparedness, including the assessment of nonlinguistic factors such as teaching ability and knowledge of American classroom norms. Not surprisingly, Xi found that the TOEFL iBT Speaking scores were more closely correlated with those local assessments that had linguistically driven rubrics and less closely correlated with local assessments with rubrics that included nonlinguistic aspects related to teaching. In addition, Xi used regression analyses to examine how well TOEFL iBT Speaking scores worked to sort potential ITAs into the different assignment categories (nonpass, provisional pass, clear pass). She found that TOEFL iBT Speaking scores were significant predictors of and accurate in classifying the ITAs' assignment categories. Xi's study provided substantial evidence for the validity of using TOEFL iBT Speaking scores for ITA screening. However, although the TOEFL iBT Speaking scores might be adequate in predicting outcomes in relation to local screening measures, Xi also acknowledged that the teaching simulation tests are more relevant for the screening of ITAs and for diagnostic purposes. Xi's study was a criterion-related validity study of the speaking section of the TOEFL iBT and did not utilize outcome variables that included student evaluations of their instructor or classroom observation assessments, and Xi specifically called for further research that uses student evaluation of ITAs in their classrooms.

Similarly, Farnsworth (2013) investigated the validity of using TOEFL iBT Speaking scores for ITA certification purposes. He argued that evidence supporting the use of TOEFL iBT Speaking for this purpose has not yet been collected and that this type of research needs to be conducted to support a validation argument, because the type of language abilities that successful instructors and teaching assistants need might be very different from the language abilities needed to be successful as a student. Using data from Xi's (2008) study, Farnsworth (2013) examined the extent to which the TOEFL iBT Speaking assessment measured the same construct as the *TOP*™ test of oral proficiency, a local ITA screening exam.

The TOP was one of the local exams investigated by Xi (2008) and does not include teaching competence in the scoring of the exam.

Utilizing factor analysis, Farnsworth (2013) concluded that the best model of speaking proficiency measured by the TOEFL iBT Speaking section was a one-factor model (oral language proficiency), whereas the factor structure for the TOP involved a higher order single factor (oral language proficiency) that loaded on three factors: pronunciation, grammar/vocabulary, and rhetorical competence (e.g., extended discourse, organization, and question handling). Farnsworth then used structural equation modeling to examine the two tests and found a higher order factor model, where the TOP second-order factor loaded onto the TOEFL speaking ability factor. Farnsworth concluded that, to a large extent, the TOEFL iBT Speaking section and the TOP were measuring the same construct (overall general oral language ability) and argued that the study provides evidence that TOEFL iBT Speaking scores can be used to make valid inferences about the oral language ability of the test takers and their ability to be ITAs. However, the author was also careful to note that this is a tentative conclusion, because the TOEFL iBT Speaking section obviously does not assess "candidates' ability to interact with a live interlocutor, manage discourse, and so on" (p. 288). Farnsworth also stated that there is a need for research that links "TOEFL speaking scores directly to real-life criterion measures of ITA language performance, such as undergraduate ratings of their ITAs' comprehensibility" (pp. 286–287), and that a very useful study would be to compare TOEFL iBT Speaking scores with the ratings of ITA performance by students, faculty, or trained observers.

Another component of the issue of ITA screening and preparation relates to the overall premise of these ITA preparation courses, which seems to be that a semester of intensive language training and instruction in effective pedagogy will prepare the potential ITAs to serve in some instructional capacity, including leading their own classes as instructors. Although each ITA program has different requirements for passing their respective ITA preparation courses, many of the courses culminate with a teaching simulation assessment where the potential ITA demonstrates his or her teaching skills and English oral proficiency. Again, the implication is that this semester of language training and instruction in pedagogy will result in increased oral proficiency for the ITAs. Indeed, the students in these classes had to enroll in the classes because their scores on the initial language proficiency screenings were too low, and passing the ITA preparation class would then imply that their language proficiency had increased to the necessary level. Although it is difficult to ascertain the information, it seems that few ITA programs do actually assess the extent to which their ITAs in preparation improve their oral proficiency over their preparation period. In other words, although tests like the TOEFL iBT are often used for initial screening of ITAs' oral proficiency, no ITA programs seem to use the TOEFL iBT to measure ITAs' improvement in oral proficiency as a result of the semester of intense language training (Farnsworth & Wagner, 2013), and thus little empirical evidence shows that the students' English proficiency actually does improve over the course of this semester of training.

Undergraduates' Evaluation and Perceptions of International Teaching Assistants

Although ITA programs provide numerous types of training (both in language and pedagogy) for ITAs to prepare them for teaching duties, the ultimate judges of ITAs' language proficiency and preparedness to teach are the undergraduates who take their classes and complete student evaluations. Numerous studies (e.g., Abraham & Plakans, 1988; Briggs & Hofer, 1991; Davis, 1991; Hinofotis & Bailey, 1981; Inglis, 1993; Rubin, 1992; Schmidgall, 2013) have investigated undergraduates' evaluations of ITAs, with the results suggesting that undergraduates' perceptions of ITAs' oral proficiency affect their perception and evaluation of the ITAs' teaching ability (often negatively, and often unfairly).

One study that investigated how ITAs' scores of oral proficiency correspond to scores of their teaching competence was conducted by Oppenheim (1998). She correlated the scores of 89 ITAs on a test of oral English proficiency with their scores on a teaching effectiveness test. The test of oral proficiency had four scales (pronunciation, worth 30% of the overall grade; grammar, 20%; fluency, 10%; and comprehensibility, 40%). The teaching effectiveness test involved a group of undergraduates evaluating the ITA's teaching performance on a videotaped simulated lesson using a teaching effectiveness scale that had four subscales: pedagogy, interpersonal skills, linguistic characteristics, and the students' familiarity with the concepts covered. Oppenheim found only a weak correlation ($r = .25$) between the ITAs' scores on the oral proficiency test and their scores on the teaching effectiveness test. The correlation was only slightly higher between the scores on the oral proficiency test and the linguistic characteristics subscale of the teaching effectiveness scale ($r = .33$).

Another study that examined how undergraduate students rated the oral proficiency of their ITA instructors was conducted by Schmidgall (2013). Using a confirmatory factor analysis approach, Schmidgall explored how both speaker- and

listener-related factors affected the undergraduate students' perceptions and judgments of the comprehensibility of the ITAs. The undergraduate students were part of the scoring panel (as naïve raters) of ITAs taking their teaching performance test (the TOP test) as the culminating project for their ITA preparation class. Schmidgall found that the students' perceptions of the ITAs' comprehensibility were affected by a number of speaker factors (oral proficiency, teaching effectiveness, teacher personality) and a number of listener factors (attitudes toward the speaker and interest in and familiarity with the topic being taught). In summary, both construct-relevant and construct-irrelevant factors influenced the students' assessments of the ITAs' comprehensibility. In Schmidgall's refined model of interactional language use and ITA speaker comprehensibility, fully 26% of the variance in the ITAs' comprehensibility was predicted by the factors related to the ITAs' teaching ability and style (teacher personality), while the speaker's pronunciation only predicted 13% of the variance. Schmidgall concluded that because the naïve student ratings of comprehensibility were influenced so dramatically by teaching factors and the teaching context, which, in theory, are not construct relevant for a test of oral proficiency, it is necessary to define the construct within a model of LSP ability (Douglas, 2000). For assessment purposes, the context in which the language is used and the strategic competence of the speaker within that context greatly influence the measurement of a test taker's ability in that particular domain.

Two somewhat dated studies have investigated directly the extent to which the undergraduates' evaluations of their ITAs' oral proficiency correlate with those same undergraduates' evaluations of the ITAs' teaching competence. Bailey (1983) had undergraduate students fill out a questionnaire rating their ITAs' oral proficiency (including pronunciation, grammar, and vocabulary). She correlated these oral proficiency scores with the ITAs' official teaching evaluation scores (the student evaluation of teaching) and found a weak to moderate correlation ($r = .46$, $p < .01$). Bailey also correlated the teaching effectiveness scores (the student evaluation of teaching) with Foreign Service Institute (FSI) oral interview scores (grammar, pronunciation, fluency, and vocabulary). She found a weak and not statistically significant correlation between the teaching effectiveness and overall FSI oral interview scores ($r = .18$, n.s.).

Inglis (1993) also investigated the correlation between ITAs' teaching competence and oral proficiency. In her study, 615 undergraduates evaluated the speech of their 16 ITAs using an eight-item questionnaire (speech evaluation scores). The teaching performance of the 16 ITAs was evaluated using official end-of-semester student evaluations. In addition, Inglis used the *SPEAK*® test to get an objective measure of the ITAs' oral proficiency. She found that the undergraduates' speech evaluation scores of the ITAs correlated moderately strongly with the teaching scores ($r = .72$, $p < .01$) and that the ITAs' SPEAK test scores correlated moderately with undergraduates' evaluations of their ITAs' oral proficiency (the speech evaluation scores; $r = .57$, $p < .05$). However, similar to Bailey's (1983) findings, the objective measure of the ITAs' oral proficiency (their SPEAK test scores) did not correlate ($r = .15$, n.s.) with their teaching scores.

Unfortunately, student evaluations of their instructors' teaching competence are often unreliable and prone to particular student biases, causing the inferences made from their results to be of dubious validity (Marsh, 1984; Rubin, 1992). Studies have found that evaluation of ITAs' teaching performance is affected by the undergraduate evaluators' native languages, grade point averages (GPAs), grade expectations, and amount of contact with nonnative speakers (Carrier et al., 1990; Jacobs & Friedman, 1988; Plakans, 1997). In addition, numerous studies (e.g., K. Brown, 1992; Nelson, 1992; Rubin, 1992, 2002; Rubin & Smith, 1990; Schmidgall, 2013; Yook & Albert, 1999) have demonstrated how students' evaluations of their instructors can be negatively influenced by their personal and cultural attitudes rather than being based solely on language and teaching ability.

Kang, Rubin, and Lindemann (2014) examined how undergraduate students' ratings of ITAs' comprehensibility and level of accentedness might be affected by the undergraduates' attitudes toward nonnative speech. They found that undergraduate students who engaged in positive intergroup contact (cooperative problem-solving exercises with ITAs) subsequently tended to rate the teaching competence and comprehensibility of ITAs more highly than did undergraduate students who did not engage in the positive intergroup contact intervention.

Obviously, the use of undergraduate students' evaluations of their (ITA) instructors is problematic, and the literature has suggested that it would be useful for researchers to use other criterion measures of ITAs' teaching performance in addition to and in conjunction with student evaluations. Nevertheless, although these undergraduate evaluations are often unreliable, biased, and problematic, they are the primary or even sole measure of teaching competence, not only for ITAs but also for all university instructors, in real-life academic contexts. Indeed, these evaluations are used for a number of high-stakes decisions, including hiring and rehiring instructors and even promotion and tenure decisions, and thus, although using student evaluations of teaching competence is problematic and even suspect, in many universities, they are

used as the ultimate arbiters of teaching competence. Those ITAs who consistently receive poor undergraduate student evaluations of their teaching are less likely to be retained as instructors than those ITAs who receive higher evaluations.

Development of Oral Proficiency

Relatively little information is available about the extent to which ITAs' oral proficiency improves over the course of their studies and employment. As stated in a previous section, there is a widespread belief that ITAs' oral proficiency improves as they live and study in an English-language environment, and many ITA programs incorporate this belief when setting cut scores on different oral proficiency assessments (Wylie & Tannenbaum, 2006; Xi, 2008). Potential ITAs who achieve the prescribed cut scores are allowed to teach, whereas those whose scores fall below the designated cut scores are not allowed to teach or have reduced teaching roles, such as individual student tutoring. Those potential ITAs who score below or near the cut scores are often required to take an ITA training course, usually focusing on oral proficiency and teaching skills. On successful completion of this course, they are then allowed to teach. The assumption here seems to be that this semester of living and studying in an English-speaking environment, in combination with the targeted training, will give the prospective ITAs the increased oral proficiency necessary to be classroom instructors.

Although this would seem to be an empirical question, the only published research on the gains in oral proficiency that ITAs might make after a semester of learning (and teaching) in a North American university appears to be Halleck and Moder (1995). They examined 15 potential ITAs who were placed into a semester-long ITA training class that focused on language skills, cultural awareness, and "compensation strategies to overcome language problems" (p. 748) as well as teaching strategies. The 15 test takers took the ITA test (similar to the TEACH test used in the current study) before the class started and again at the end of the class. Halleck and Moder found that, as a group, the 15 ITAs scored higher on the end-of-class ITA test in all three categories (presentation language skills, interactive language skills, and teaching skills). However, Halleck and Moder concluded that the higher ability test takers (those seven ITAs who scored highest on the initial ITA test) seemed to benefit more from the ITA preparation class than did the seven ITAs who scored lowest on the initial ITA test, especially in relation to teaching skills. They argued that this demonstrated that there seems to be a minimum threshold of language ability below which ITA training might not be all that useful for potential ITAs. Although Halleck and Moder's study is very informative in investigating the improvement in language proficiency over the course of a semester-long ITA preparation class, empirical studies have not directly investigated ITA oral language improvement through measurements on large-scale, standardized proficiency assessments like the TOEFL iBT or IELTS.

However, numerous studies have investigated gains in proficiency by second language (L2) learners after a semester in an immersion environment. Often these studies (Freed, 1995; Segalowitz & Freed, 2004; Towell, Hawkins, & Bazergui, 1996) have examined the extent to which American foreign language learners in a study abroad context improve in their oral proficiency. Overall, these studies suggest that these types of learners make quite limited gains after one or two semesters of study abroad. Freed (1998) performed a review of the study abroad studies and found that most reported only limited gains in oral proficiency for the L2 learners in the study abroad context. More importantly, she found that there was a great amount of individual variation in oral proficiency gains and that the gains were primarily in greater ease and confidence in speaking (i.e., greater abundance of speech, faster speech rate, and fewer disfluent pauses). Freed also concluded that the research is lacking about the changes in structural accuracy and suggested that, for more advanced learners, there is little change in grammatical accuracy.

For L2 learners of English, the results overall are similar. Derwing, Munro, and Thomson (2008) investigated the development of two groups of ESL learners' fluency and comprehensibility over a 22-month period and found that one group (Slavic language speakers) demonstrated only a very small increase in fluency and comprehensibility, whereas the second group's (Chinese speakers) fluency and comprehensibility did not improve over the same period. The authors attributed this lack of oral proficiency development in part to the learners' lack of exposure to English outside of their ESL classrooms. Similarly, O'Loughlin and Arkoudis (2009) investigated the IELTS score gains made by nonnative speakers of English after a course of study at an Australian university and found (similar to Freed's findings reported earlier) that there was a great deal of individual variation in the amount of language gain but, as a group, the participants' gains were surprisingly limited, even after a number of years in English-speaking academic environments. The participants improved the most in listening and reading, with much smaller gains in writing and speaking. Elder and O'Loughlin (2003) investigated the band score gains on the IELTS after a 10- to 12-week intensive English-language study program in Australia

and New Zealand. The overall findings were that the students made very limited gains in their English proficiency. Elder and O'Loughlin found that listening was the skill with the greatest gain, whereas reading had the lowest level of gain.

Ling, Powers, and Adler (2014) investigated the development of proficiency of two different groups of English-language learners: a group of Chinese high school English as a foreign language (EFL) learners over a 9-month period and a group of ESL learners in an intensive English-language course in the United States over a 6-month period. Similar to the present study, the participants took the TOEFL iBT the beginning of the study and again at the end of the study to investigate how much the proficiency of the participants had increased. The researchers found that both groups made substantial gains as measured by the effect sizes of the TOEFL iBT score gains. The group of Chinese high school EFL learners improved substantially on the reading and listening sections of the TOEFL iBT but less on the writing and speaking sections. In contrast, the group of ESL learners improved substantially on the speaking and writing sections, but less so on the reading and listening sections. According to Ling et al., these results suggest that the immersive environment for the ESL students might be more useful in improving students' productive (speaking and writing) abilities and have less of an impact on their receptive abilities, whereas the EFL instruction in China might focus more on developing learners' receptive skills. Ling et al. also investigated the study habits of the participants and found (perhaps not surprisingly) that those learners who spent greater amounts of time studying outside the classroom tended to have higher test scores and greater gains in the scores in general. The authors also concluded that the TOEFL iBT is able to measure the improvement that learners make in their English proficiency.

For most of the ESL studies cited earlier, the conditions would be considered ideal for optimal language learning; that is, the learners were in an immersion environment, surrounded by the target language. These learners lived in a country where the target language was spoken and were also attending college or university and taking classes in the target language (except for the subjects of Derwing et al.'s, 2008, study), and thus the learners had the opportunity to learn in the classroom as well as naturalistically. Ranta and Meckelborg (2013) argued that "living in an environment where the target language is used daily is expected to provide abundant opportunities for learners to engage in target-language use, resulting in an enhanced ability to understand and produce fluent, colloquial speech" (p. 2). Nevertheless, the learners in the studies cited (other than the Ling et al., 2014, study) made, at best, only limited gains in proficiency. In trying to interpret these limited gains, a common conclusion by the researchers was that even though the participants were in an immersion setting, they still had surprisingly little interaction with target language speakers outside of the classroom. Focusing on undergraduate and graduate L2 learners in English-speaking countries, a number of studies (e.g., Cheng & Fox, 2008; Myles & Cheng, 2003; Ranta & Meckelborg, 2013; Trice, 2004) have found that these L2 learners often have limited interaction with native English speakers and, consequently, less exposure to English than would be expected. Freed (1998) concluded that study abroad learners' "interactions with native speakers may be far less intense and frequent than was once assumed" (p. 50). Even the results of the Ling et al. (2014) study, which found the ESL learners making substantial gains in their English proficiency over the course of their study, suggest that the learners who engaged in the most out-of-classroom study seemed to improve the most, but the authors also concluded that engagement does not guarantee success, because "simply associating with native English-speaking peers (or faculty members who are proficient English users) does not make students more efficient English learners" (p. 14).

This issue of the amount of interaction international students have on campus with native English speakers was the specific focus of Ranta and Meckelborg's (2013) study. They conducted a 6-month longitudinal study of 17 Chinese graduate students at a Canadian university (and some of these participants were ITAs). Whereas previous studies of immersion learners' amount of interaction with English speakers and exposure to English relied on diary reports and questionnaires, the researchers in this study used a computerized language log to quantify the participants' amount of interaction in and exposure to English. This system led to a much more reliable quantification of the participants' interaction and exposure to English than previous studies. Not surprisingly, the study found that the amount of English used by the participants varied according to the different activities in which they were engaged (i.e., English for academic work, attending class, teaching assistant/research assistant activities; more Mandarin use for social interaction, recreation, and daily life). An interesting finding is that the amount of daily use of English by the participants actually decreased slightly over the 6-month period. The study also found a great deal of variation in the amount of English used among the 17 participants in different activities, including in the time spent watching movies or TV in English, reading for pleasure in English, and total oral interactions in English, and fully half of the participants reported doing no reading for pleasure in English. The authors concluded that although the participants spent more hours each day

using English than their first language (L1), the quality of the exposure was often low, although this varied greatly by individual.

A final issue related to the development of ITAs' oral proficiency that must be addressed is the difficulty in reliably assessing the development of oral proficiency over a relatively small time frame. Zhang (2008) investigated the reliability and validity of the TOEFL iBT by performing repeater analyses on the scores of 12,385 test takers who took the TOEFL iBT two times within a 1-month period in 2007. Zhang found that the increase in scores from the first test to the second test was very small for the overall test and that the gains in the scores on the speaking, listening, and writing components of the TOEFL iBT were smaller than the gains on the reading section of the test. These results suggest that the TOEFL iBT Speaking and Listening sections are reliable measures of oral proficiency. Indeed, Ling et al. (2014) used the TOEFL iBT to assess ESL learner development over a 6-month period and EFL learner development over a 9-month period and concluded that the TOEFL iBT "can capture changes (or improvement) in English-language skills as a result of learning and instruction" (p. 13).

In conclusion, this review of the literature makes clear a number of points in relation to the use of TOEFL iBT scores for ITA screening and informs the research focus of the present study. First, although there is much theoretical justification for the use of both TOEFL iBT Speaking and Listening scores for ITA screening purposes, it seems that all the ITA programs reported in the literature that do use the TOEFL iBT scores for screening purposes rely on the TOEFL iBT Speaking scores exclusively. Second, although a number of useful and informative validation studies have been conducted on the use of TOEFL iBT scores for ITA placement purposes, there is a need for the inclusion of outcome measures of both language and teaching competence in these validation studies. Related to this idea of including outcome measures of teaching competence, although formal student evaluations of instructional competence are a necessary outcome measure to include in these studies (because they are so prevalent and because they are used to make high-stakes decisions), these formalized and institutional student evaluations are problematic for a number of reasons. Thus, alternative measures of language and teaching competence, including focused student evaluations, as well as objective observers' evaluations of teaching and language competence, would be useful outcome measures to incorporate into the validation studies. Finally, there is a surprisingly significant gap in the literature regarding the oral proficiency development of ITAs over the course of a semester of training and teaching. Although many ITA programs seem to take for granted that potential ITAs will show marked increases in oral proficiency from taking an ITA class and living and studying on a university campus, almost no research has actually tried to quantify the gains in oral proficiency that ITAs make.

Research Questions

Based on this review of the literature, the following research questions were formulated to guide the current study:

1. To what extent do TOEFL iBT Speaking and Listening scores correspond to other criterion measures of ITA teaching preparedness used at Temple University?
2. To what extent do TOEFL iBT Speaking and Listening scores correspond to measures of ITAs' classroom teaching performance?
3. What gains in oral proficiency (as measured by TOEFL iBT Speaking and Listening scores) do ITAs make after a semester of ITA training and/or teaching?

Method

This study has two major components, and two different sets of data were collected. The first set of data consists of existing data that were compiled from Temple University databases. Data for this part of the study were from the Fall 2006 semester through the Spring 2012 semester. The second set of data consists of new data collected specifically for this study, from the Fall 2009, Spring 2010, Fall 2010, Spring 2011, Fall 2011, and Spring 2012 semesters.

Participants

In any given year, approximately 30 prospective ITAs come to Temple University. These ITAs are graduate students (usually in doctoral programs, although some are in master's programs) generally ranging in age from 21 to 30 years. They come from a variety of countries, including China (approximately 40%), India (approximately 20%), Iran (approximately 10%),

Table 1 English Proficiency Requirements for International Teaching Assistants (ITAs) at Temple University

TOEFL iBT Speaking scores	SPEAK test scores	ITA 5501	TEACH test
≥28 or above	Not required	Not required	Not required
<28	≥50 (pass)	Not required	Not required
<28	45–49 (restricted pass)	Required but still able to teach	Must pass to assume full instructional duties
<28	<45 (nonpass)	Required, not able to teach	Must pass to assume full instructional duties

and Nepal, South Korea, Serbia, Turkey, Russia, Pakistan, Sri Lanka, and other countries (approximately 30% total). The majority of the ITAs are in the College of Science and Technology or the College of Engineering, but others are in the Colleges of Music and Dance, Education, Liberal Arts, Communications and Theater, and Business and Management.

As mandated by Pennsylvania law, to be instructors at the college level, the prospective ITAs must demonstrate their proficiency in English. From 2006 to the present, the Temple University ITA program has used TOEFL iBT Speaking scores (or IELTS scores) as an initial screening measure of ITAs' oral proficiency. Those with a TOEFL iBT Speaking score of 28 or higher may assume instructional duties without restriction. ITAs with lower scores are required to take the locally administered SPEAK test. On the basis of their scores on the SPEAK test, they are assigned to one of three groups: (a) A score of 50 or higher is a pass (unrestricted instructional duties), (b) a score from 45 to 49 is a restricted pass (may assume instructional duties for one semester while simultaneously enrolled in Temple's ITA training course, ITA 5501), or (c) a score less than 45 is a nonpass (need to pass ITA 5501 before assuming instructional duties). To pass ITA 5501, ITAs must have a grade assigned by the ITA 5501 instructor of at least 80% based on attendance and participation, homework, and small projects. They must also pass a teaching simulation test (TEACH) administered at the end of the class and judged by a panel consisting of ESL teachers in the Intensive English Language Program (IELP) at Temple. This information is shown in Table 1.

From Fall 2008 through Fall 2012, approximately 150 new ITAs came to Temple. Approximately 10% of the prospective ITAs who applied to Temple University scored at least 28 on the speaking section of the TOEFL iBT and thus were able to assume instructional duties without taking any local screening tests. Approximately 50% of the incoming ITA candidates scored less than 28 on the speaking section of the TOEFL iBT but scored at least 50 on the SPEAK test and thus were able to assume full instructional duties with no other requirements. Approximately 25% of incoming ITAs scored between 45 and 49 on the SPEAK test and thus were able to assume instructional duties but were also required to simultaneously take ITA 5501. Approximately 15% of incoming ITAs scored less than 45 on the SPEAK test and thus could not assume instructional duties until they had successfully completed ITA 5501 (and passed the TEACH test).

Participants for Research Question 1

Research Question 1 investigated the correlations between TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK test scores. Data from a total of 198 ITAs were included in this component of the study, and from this overall sample, subsamples were used for subsequent components of the study (described later). This overall sample of 198 ITA participants is representative of the ITA population at Temple, and the demographic information of this sample is given here. Of the 198 ITA participants, 121 were male (61.1%) and 77 were female (38.9%). They came from 29 different countries, with 79 (39.9%) from the People's Republic of China; 37 (18.7%) from India; 19 from Iran (9.6%); eight from Nepal (4%); eight from South Korea (4%); six from Serbia (3%); five from Turkey (3%); five from Russia (3%); three each from Italy, Japan, and Taiwan; two each from Colombia, Germany, Nigeria, and Ukraine; and one each from Bangladesh, Bulgaria, Canada, Chile, Egypt, Indonesia, Jordan, Kazakhstan, Lebanon, Morocco, Paraguay, the Philippines, Sri Lanka, and Uganda. The ITAs reported 23 different first languages, with the most common being Chinese (41.4%), Telugu (7.1%), Farsi (9.6%), Hindi (6.1%), Bengali (5.1%), Korean (4%), Nepali (4%), Russian (4%), and Serbian (3%). Other L1s reported included Arabic, Bulgarian, Filipino, French, German, Ibo, Indonesian, Italian, Kurdish, Luganda, Marathi, Sinhala, and Spanish. Almost 60% ($n = 117$) of the ITAs in the study were in the College of Science and Technology, 20.7% ($n = 41$) came from the College of Engineering, and 7.1% ($n = 14$) came from the College of Liberal Arts. The remaining 28 ITAs came from the College of Music and Dance, College of Education, College of Sociology, School of Business, School of Communication and Theater, School of Pharmacy, School of Tourism and Hospitality Management, and School of Art. The ITAs were graduate students in 32 different academic departments, with the most common being the Department of Computer and

Information Systems ($n = 48$, or 24.2%), Department of Chemistry ($n = 39$, or 19.7%), Department of Mechanical Engineering ($n = 18$, or 9.1%), Department of Physics ($n = 13$, or 6.6%), Department of Electrical and Computer Engineering ($n = 13$, or 6.6%), Department of Civil and Environmental Engineering ($n = 10$, or 5.1%), Department of Mathematics ($n = 8$, or 4%), and Department of Biology ($n = 8$, or 4.0%).

From this sample of 198 ITAs, a subsample was used to investigate the correlations between TOEFL iBT Speaking, TOEFL iBT Listening, the SPEAK test, and TEACH test scores. It was necessary to use only a subsample, because of the 198 ITAs in the sample, only 89 ITAs also had TEACH test scores (described in more detail in the "Results" section). The background characteristics of this subsample of 89 ITAs are similar to the overall 198 ITA sample. Of these 89 ITA participants, 48 (65.2%) were male and 31 (34.8%) were female. They came from 15 different countries, with the most common native countries being China ($n = 45$), India ($n = 13$), Iran ($n = 9$), and South Korea ($n = 6$). These 89 ITAs reported 15 different L1s, with the most common being Chinese ($n = 48$), Farsi ($n = 9$), Telugu ($n = 8$), and Korean ($n = 6$). More than two-thirds of this sample of ITAs came from the College of Science and Technology ($n = 61$, or 68.5%); 18 (20.2%) came from the College of Engineering.

Participants for Research Question 2

Research Question 2 used existing data from all the ITAs from 2006 to 2012 who taught classes in their first semester at Temple, including those ITAs who scored at least 28 on the TOEFL iBT as well as those who scored lower than 28 but received a score of 45 or higher on the SPEAK test. This resulted in 128 ITAs for which there were TOEFL iBT Listening scores, TOEFL iBT Speaking scores, SPEAK test scores, and official student evaluations of their instructors, referred to here as student feedback form (SFF) scores. The background characteristics of this subsample of 128 ITAs were similar to the overall 198 ITA sample described earlier. Of these 128 ITA participants, 81 (63.3%) were male and 47 (36.7%) were female. They came from 25 different countries, with the most common native countries being China ($n = 54$), India ($n = 25$), Iran ($n = 7$), and Nepal ($n = 6$). These 128 ITAs reported 22 different L1s, with the most common being Chinese ($n = 56$), Bengali ($n = 10$), Telugu ($n = 9$), Nepali ($n = 6$), and Hindi ($n = 6$). More than two-thirds of this sample of ITAs came from the College of Science and Technology ($n = 90$, or 70.3%); 18 (14.1%) came from the College of Engineering.

The second set of analyses for Research Question 2 used newly collected data, including the undergraduate students' assessments of their ITAs' language proficiency and teaching competence as well as the observers' assessments of the ITAs' language proficiency and teaching competence. Thirty-three ITA participants had complete sets of data for these analyses. These 33 ITAs were from the overall sample of 198 ITAs described earlier. Of these 33 ITA participants, 23 (69.7%) were male and 10 (30.3%) were female. They came from 10 different countries: China ($n = 19$); Turkey ($n = 4$); Serbia and India ($n = 2$ each); and Germany, Iran, Italy, Japan, Nepal, and Sri Lanka ($n = 1$ each). Twenty-seven of the ITAs in this sample came from the College of Science and Technology (82%), and the remaining six ITAs (12%) came from other schools and colleges at the university.

Participants for Research Question 3

The analyses for Research Question 3 necessitated that participants take the TOEFL iBT Speaking and Listening both at the beginning and at the end of their first semester as ITAs at Temple. Of these 84 ITA participants, 50 (59.5%) were male and 34 (40.5%) were female. They represented 16 different countries, with the most common native countries being China ($n = 45$), India ($n = 16$), and Iran ($n = 10$). These 84 ITAs reported 16 different L1s, with the most common being Chinese ($n = 43$), Farsi ($n = 10$), and Hindi ($n = 4$). The vast majority of this sample of ITAs came from the College of Science and Technology ($n = 29$, or 94.0%); 5 (6.0%) participants came from four other colleges and schools at the university.

Instruments

TOEFL iBT Speaking and Listening

The TOEFL iBT was designed to assess the communicative language proficiency of test takers (Butler et al., 2000; Chapelle, Enright, & Jamieson, 2008). For this study, two versions of the TOEFL iBT were used. These versions were a result of the TOEFL iBT Research Form Creator, which is made available to researchers conducting research on the TOEFL iBT,

and they were installed on 16 computers in a computer lab at Temple University. These two versions of the TOEFL iBT are virtually identical to the operational TOEFL iBT, including the interface and delivery platform (except that only the speaking and listening sections were administered).

The speaking component of the TOEFL iBT includes six tasks, with an academic TLU domain: Two of the tasks require the test taker to talk about everyday topics; two involve typical college campus situations; and two involve content typical of college courses. These tasks also have different formats: Two require the test taker to respond orally to a written prompt; two require the test taker to listen and respond to an oral prompt; and two require the test taker to respond to both written and spoken prompts (Wylie & Tannenbaum, 2006). The speaking tasks are scored using a holistic rubric that has a 5-point scale ranging from 0 to 4 with four categories: general description, delivery, language use, and topic development (for the rubric, see ETS, 2014).

Bridgeman, Powers, Stone, and Mollaun (2012) provided validity evidence that the TOEFL iBT Speaking does indeed measure a speaker's communicative competence. They found very strong correlations between the TOEFL iBT Speaking scores from 184 test takers and undergraduate students' comprehension ratings of the speech samples from TOEFL iBT Speaking forms from those same 184 test takers. Biber and Gray (2013) provided evidence for the validity argument of the TOEFL iBT when they performed a corpus study of TOEFL iBT test takers' production and demonstrated that the test elicits features of real-life language use. They argued that the vocabulary, discourse, collocations, and lexicogrammatical features of the written responses of the TOEFL iBT were similar to those vocabulary, discourse, collocations, and lexicogrammatical features typical of academic written language, while the vocabulary, discourse, collocations, and lexicogrammatical features found in the test takers' spoken responses were similar to those features found in the spoken corpora; they also argued that test takers' variation across written and spoken texts was similar to the variation in real-life language use.

In contrast to the studies just described, Brooks and Swain's (2014) study provided less clear-cut evidence supporting the validity argument for the TOEFL iBT Speaking. Brooks and Swain compared the performance of 30 L2 graduate students on TOEFL iBT Speaking tasks and in real-life academic speaking tasks. They examined the grammatical, discursive, and lexical features found in the two sets of speaking tasks and found that although some of the features were overlapping (connectives, nominalization, vocabulary), other features were distinct to one of the two performances (e.g., grammatical complexity and inaccuracy, use of informal language, speech organizers). Brooks and Swain argued that this provided only limited evidence supporting a validity argument for the use of the TOEFL iBT Speaking.

Sawaki and Sinharay (2013) investigated a number of qualities of the TOEFL iBT, including the dimensionality of the test and the usefulness (and reliability) of reporting the different subscores. Their findings based on exploratory factor analyses were similar to the findings of a number of previous studies that the TOEFL iBT seemed to have two correlated factors: a speaking factor and a reading/listening/writing factor. However, Sawaki and Sinharay's analyses involving confirmatory factor analysis resulted in a slightly different conclusion, which was a four-factor model (listening, speaking, reading, and writing). The four factors were highly correlated, although the speaking factor "was relatively more distinct from the others" (p. 80).

The listening component of the TOEFL iBT includes a total of six to nine listening tasks. These tasks include four to six lecture texts, some of which include classroom discussion. Each of these texts is 3 – 5 minutes in length and includes six comprehension questions. There are also two or three tasks with conversation texts. Each of these texts is approximately 3 minutes in length, with five comprehension questions (ETS, 2005).

Sawaki and Nissan (2009) conducted a criterion-related validity study of the listening section of the TOEFL iBT. In their study, Sawaki and Nissan surveyed 145 undergraduate and graduate students at four North American universities, asking them about the types of listening tasks that were required in their university classes and course assignments. They then used the results of this survey to create a listening assessment that used an academic lecture text that included the types of listening tasks that the survey respondents reported having to do in university classes and course assignments. Sawaki and Nissan correlated the scores for the 184 participants on this listening assessment with their TOEFL iBT Listening scores. They found a moderate correlation ($r = .64$ for the observed correlation; they also reported the disattenuated correlation, $r = .70$) between the TOEFL iBT Listening scores and the criterion measure and concluded that these results provided criterion-related evidence of the validity of using TOEFL iBT scores as a predictor of academic listening ability.

Chapelle (2008), and Xi, Bridgeman, and Wendler (2013) have argued that for academic English proficiency tests such as the TOEFL iBT (which can be considered a test for specific purposes), it is necessary to analyze the TLU domain of

universities for the types of language used there. As stated earlier, a number of studies (e.g., Biber & Gray, 2013; Bridgeman et al., 2012) have been conducted providing empirical evidence for the validity argument in support of the TOEFL iBT Speaking. However, there does not seem to be the same amount of research analyzing the TLU domain of academic listening as part of the validity argument of the TOEFL iBT Listening section (other than Sawaki & Nissan, 2009). Indeed, Wagner (2012, 2014a, 2014b) reviewed large-scale English proficiency tests used for North American university admissions (including the TOEFL iBT), and Wagner and Wagner (2016) did the same with large-scale English proficiency tests used in Asia. They noted that the spoken texts used on the listening sections of the tests lack many of the characteristics of real-world spoken language, including connected speech, hesitation phenomena, spoken grammatical norms, and the nonverbal information transmitted by the speaker. In addition, while these studies (Biber & Gray, 2013; Bridgeman et al., 2012; Sawaki & Nissan, 2009) have provided validity evidence for the TOEFL iBT Speaking and Listening sections, they have focused on using the TOEFL iBT as a measure of test takers' ability to perform as a student at a North American university. But the use of the TOEFL iBT as a screening measure for instructors would seem to be (at least in part) targeting a different language use domain (Farnsworth, 2013; Farnsworth & Wagner, 2013). Although there is substantial evidence supporting the use of TOEFL iBT test scores as a measure of a student's ability and readiness for study in an English-medium university, a validity argument needs to be made supporting the use of TOEFL iBT test scores as a measure of an instructor's ability to teach in an English-medium university.

All prospective ITAs who apply to Temple University are required to submit TOEFL test or IELTS scores as part of the admission process. The ITAs' TOEFL iBT Speaking and Listening scores that were submitted to Temple as part of their applications for admission were collected from the Measurement and Research Center at Temple for use in this study. In addition, 84 new ITAs from 2009 to 2012 participated in this study and took a research version (based on the TOEFL iBT Research Form Creator) of the TOEFL iBT two times: at the beginning and at the end of their first semester of ITA activity at Temple. These administrations of the TOEFL iBT were conducted at Temple by the researcher, and only the speaking and listening sections were administered. Each speaking sample was rated by two trained and certified ETS raters. ETS raters go through an extensive training and norming process. Once they are certified as raters, they are then able to rate the speaking test samples. For operational TOEFL iBT tests, each speech sample is rated by at least one rater, and a subsample is rated by two raters to assure rating reliability. For this study, however, it was decided to have all of the speaking samples rated by two raters. A pool of 16 certified ETS raters rated the speech samples and were paid for their ratings.

The SPEAK Test

The SPEAK test of oral proficiency was created by ETS. It consists of 12 speaking tasks in which the test takers speak their responses into an audio recorder. Two raters score the 12 speech samples using a rubric. These raters are certified to score the samples and are ESL teachers in the IELP at Temple, in conjunction with the Temple ITA Office. The holistic rubric advises raters to consider functional competence, sociolinguistic competence, discourse competence, and linguistic competence when assigning scores using band descriptors that range from 20 to 60 in 10-point increments. The scores are averaged for the 12 tasks across both raters, and thus scores can range from 20 to 60. The SPEAK test is composed of retired versions of the Test of Spoken English (TSE). Powers, Schedl, Wilson Leung, and Butler (1999) performed a validation study of the TSE in which undergraduate students (predominantly L1 English speakers) rated the spoken texts of TSE test takers. In addition, a listening comprehension test was composed using the spoken responses of the TSE test takers. Analyses found strong correlations between the undergraduate ratings of the spoken texts, the official TSE scores, and the listening comprehension scores, providing validity evidence supporting the use of TSE scores (and, subsequently, the SPEAK test scores) for ITA screening purposes.

As described earlier, approximately half of ITA candidates scored less than 28 on the speaking section of the TOEFL iBT yet scored at least 50 on the SPEAK test. This might be problematic and might be indicative of security concerns regarding the SPEAK test. In 2006, ETS stopped supporting the SPEAK test; no new versions have been created since then, and only a limited number of versions of the SPEAK test are used. There is a concern that this lack of security could compromise the results of the SPEAK test. For these reasons, the ITA program at Temple has expressed a desire to use a different test for ITA screening purposes but has not yet moved to do so. Another concern is that although the raters are certified to score the SPEAK test, there is no ongoing effort to assess the reliability of the SPEAK test raters.

The TEACH Test

The ITA TEACH test is administered by the Temple ITA Office to all ITA students who have taken ITA 5501. The curriculum for ITA 5501 was designed specifically to teach ITAs the skills needed to teach classes, focusing on both language skills and teaching and interactional skills. The TEACH test is the culminating assessment for ITA 5501 and is used to assess the ITAs' mastery of the ITA 5501 curriculum.

For the TEACH test, the ITA teaches a simulated mini-lesson to a small group of raters meant to simulate an undergraduate class. The raters are all ESL instructors in the IELP. The analytic rubric for the TEACH test was developed by Temple's ITA program in the early 2000s and was informed by Bailey (1985), Plakans and Abraham (1990), Smith, Myers, and Burkhalter (1992), and Briggs (1994). The scoring rubric has two major categories: presentation language skills and teaching skills/interactional skills. The test taker's presentation language skills are assessed using a 4-point scale ranging from 1 (*inadequate*) to 4 (*excellent*) in five separate categories: (a) comprehensibility, (b) accent and pronunciation, (c) listening comprehension, (d) fluency, and (e) grammar and word choice. The test taker's teaching skills/interactional skills are also assessed in three separate categories on the same 4-point scale: (a) lesson organization and implementation, (b) relevance of content and development of content, and (c) interaction with students and nonverbal communication. The rater is also provided space to make written comments regarding the examinee's overall performance, which are used as formative feedback to the examinee. As described subsequently, the TEACH test scores were used in the analyses using the two parts of the test separately (TEACH Language and TEACH Teaching) and combined (TEACH Composite). The rubric is provided in Appendix A.

Student Feedback Forms

All classes at Temple include student course and teaching evaluations commonly known as the "Student Feedback Forms" (SFFs). This is similar to the formal systems that most universities have in place to evaluate instructional competence. These SFFs are created, maintained, administered, and collected by the Measurement and Research Center at Temple. From 2009 to 2013, SFFs were filled out by students at the end of the semester, in the penultimate or final class session, after the instructor left the room.[1] The SFFs consist of two sets of items to which students respond using a 5-point Likert-type scale ranging from 5 (*strongly agree*) to 1 (*strongly disagree*). The first set contains eight items evaluating the instructor's performance. The second set contains three items evaluating the course itself. Different colleges or schools might have slightly different versions of the SFF, but the format of the forms across colleges is identical, and most of the questions are the same across the different forms. The standard SFF form with the usual 11 items can be seen in Appendix B. However, not all of the SFF items were relevant for this study, so only the results of three SFF items were analyzed: (a) "The instructor was well organized and prepared for class," (b) "The instructor taught this course well," and (c) "I learned a great deal in this course."

Only these three items were included as the measure of teaching competence for a number of reasons. First, the SFF forms changed in 2010, and old items were deleted and new items were created; however, these three items remained the same on the new forms. Second, different versions of the SFF are given in different colleges and schools, but these three items are consistent across all forms. Finally, these three items are considered the most relevant for decisions including faculty and instructor hiring, rehiring, and promotion and tenure (J. Degnan, Director of the Temple Measurement and Research Center, personal communication, July 9, 2010), and thus it was decided to focus on these three items as measures of teaching competence. As was noted in the review of the literature, student evaluations of their instructors are prone to a number of biases, and using these as a measure of teaching competence is problematic. Nevertheless, these are the measures that Temple University uses (and are similar to measures used by the vast majority of colleges and universities in the United States) for a number of purposes, including high-stakes decisions regarding retention of instructors and promotion; thus they are used here.

Undergraduate Student Evaluation of International Teaching Assistant Teaching Competence

The official undergraduate SFF items employed by Temple were created to allow students to assess the quality of the class and an instructor's teaching. However, these evaluations are very broad and not focused on an ITA context. Therefore, it was decided to create an alternative measure of teaching competence and oral proficiency to use in this study, in addition

Table 2 Reliability Coefficients for the Subscales of the Student Questionnaires

Scale	Subscale	Reliability α (M)	SD
Teaching competence	Interaction with students	.71	.16
	Knowledge of American classroom cultural norms	.81	.10
	Ability to communicate content information	.83	.12
Oral proficiency	Fluency	.75	.13
	Comprehensibility and accent	.78	.15
	Grammatical and vocabulary accuracy	.81	.11
	Ability to understand students' speech	.87	.10

to the SFFs. This assessment was a questionnaire that consisted of two main scales. The first scale was designed to allow the undergraduate students in a class to evaluate their ITA's teacher–student interaction skills (the second scale was designed to allow the students to evaluate their ITA's oral proficiency and is described in the next section). This questionnaire was designed to evaluate the same abilities that are assessed in the classroom observations (described later) and are the same teaching abilities that are the focus of the curriculum of the ITA 5501 class. Students had to respond to a series of statements using a 5-point Likert-type scale ranging from 5 (*strongly agree*) to 1 (*strongly disagree*). Three subscales[2] were meant to assess the ITA's teacher–student interaction skills: (a) three items measuring interaction with students, (b) three items measuring the instructor's knowledge of American classroom cultural norms, and (c) two items measuring the instructor's ability to communicate content information.

Undergraduate Student Evaluation of International Teaching Assistant Oral Proficiency

The second part of the questionnaire was designed to allow undergraduate students to assess their ITAs' oral communication skills and to evaluate the same oral communication skills that are assessed in the classroom observations (described later) and the same language abilities that are the focus of the curriculum of the ITA 5501 class. Four subscales were used: (a) two items measuring the instructor's fluency, (b) three items measuring comprehensibility and accent, (c) two items measuring the instructor's grammatical and vocabulary accuracy, and (d) two items measuring the instructor's ability to understand students' speech.

The questionnaire includes two additional scales. The first scale consists of two dichotomous items used to measure the student's determination of whether the instructor's level of oral proficiency was sufficient to be a teacher. The second additional scale consists of three items designed to measure the student evaluator's familiarity and experience with nonnative speakers of English. These 22 items can be seen in Appendix C. This questionnaire was administered two times to each class: during the first and second observations of the ITA instructor.

This 22-item questionnaire was developed based on procedures suggested by Wagner (2010b) for questionnaire development. It was refined with feedback provided by ITAs who were students in ITA 5501 prior to this study and with the ITA 5501 instructor. In addition, the individual items were examined to ensure that they were reliably measuring what they were designed to measure. The internal consistency reliability of the set of items in each of the four subscales that were designed to measure the instructor's oral language proficiency was estimated. Similarly, the internal consistency reliability of the set of items in each of the three subscales that were designed to measure teacher–student interaction skills was estimated. In addition, the item-total correlation for each of the items was examined to ensure that each item was reliably measuring the same construct. Based on this piloting procedure, the final 22-item questionnaire used here was adopted. In the final analysis, 53 different classes consisting of 917 undergraduate students completed the questionnaire. The internal consistency reliability was estimated for each subscale of the questionnaire, and this coefficient (alpha) was averaged across the 53 different classes. The reliability ranged from a low of $\alpha = .71$ for the interaction with students subscale on the teaching competence scale, to a high of .87 for the ability to understand students' speech subscale on the oral proficiency scale. The reliability coefficient for each of the subscales is given in Table 2.

Classroom Observation Rubric

In their first semester of teaching, the ITA participants were observed and evaluated in their classrooms. Each ITA was observed two times, by two different people (the research assistants or the researcher). There were three research assistants,

Table 3 Summary of the Instruments Used in the Study

Instrument	Administered by	When administered	Key characteristics
TOEFL iBT Speaking and Listening	Researcher/research assistant in computer lab	Beginning and end of semester of first semester of ITA activity	Research version of TOEFL iBT; scored by ETS-certified raters
SPEAK Test	ITA Office staff	When potential ITAs first come to campus	Scored by certified raters (ESL instructors and ITA coordinator)
TEACH test	ITA 5501 instructor	Culminating assessment at the end of the class	Scored by ESL teachers who are also simulated class members; five subscales of oral proficiency, three subscales of teaching skills
Student feedback forms	Instructor of the class/lab	Penultimate or final class session	Developed and scored by the university's Measurement and Research Center
Undergraduate student evaluation of ITA teaching competence	Researcher/research assistant; completed by students in the ITAs' classes	Two times during the semester	Eight Likert items on 22-item questionnaire measuring three subscales of teaching competence
Undergraduate student evaluation of ITA oral proficiency	Researcher/research assistant; completed by students in the ITAs' classes	Two times during the semester	11 Likert items on 22-item questionnaire measuring four subscales of oral proficiency
Classroom observation of ITA teaching competence	Researcher or research assistant	Two times during the semester	Measures four subscales of teaching competence
Classroom observation of ITA oral proficiency	Researcher or research assistant	Two times during the semester	Measures four subscales of oral proficiency

Note. ESL = English as a second language; ITA = international teaching assistant.

all of whom were doctoral students in the Teaching English to Speakers of Other Languages (TESOL) program at Temple. Two were former ITA 5501 teachers, and the third was familiar with the ITA 5501 curriculum.

These observers used an analytic rubric with a 5-point scale to evaluate the oral proficiency and teaching competence of the ITAs. This rubric was originally created using a slightly modified version of the analytic rubric used in the TEACH test evaluation (see earlier). The rubric comprises two components: measurement of the instructor's oral skills and measurement of the instructor's teaching and interactional skills. Four categories of the instructor's oral proficiency are assessed with the rubric: (a) fluency, (b) comprehensibility and accent, (c) grammatical and vocabulary accuracy, and (d) listening comprehension. Four categories of the instructor's teaching and interactional skills were measured: (a) lesson organization and preparedness, (b) teacher–student interaction and knowledge of classroom culture, (c) delivery strategies and nonverbal communication, and (d) communication of content information.

The rubrics and observation procedure were piloted by the researcher and two of the research assistants in the Spring 2009 and Summer 2009 semesters by observing and rating a number of ITA-led classes. Based on this pilot testing, the rubric was revised to ensure reliability, and the observation procedure was finalized. Prior to the Fall 2009 semester, an extensive observation rating and rubric norming session was conducted by the researcher with the research assistants. This classroom observation rubric can be seen in Appendix D.

Table 3 provides a summary of the seven instruments used in the study, highlighting key characteristics and features of the instruments.

Data Collection Procedures

This study was approved by Temple's Institutional Review Board prior to any collection of data. All the participants in the study were informed of the purpose and rationale for the study, and all signed the necessary informed consent forms.

The first set of data used in the study involved preexisting data. The Measurement and Research Center at Temple provided the admissions TOEFL iBT Speaking and Listening scores and SFF score data. The ITA Office provided the SPEAK test and TEACH test data. The second set of data was collected specifically for this study and is described in more detail here.

Evaluation of International Teaching Assistants' Oral Proficiency and Teaching Competence

Each semester over the course of the data collection for this study, all of the ITAs who were teaching an undergraduate class for the first time at Temple were asked to participate in the study. Those who agreed were observed and evaluated by the researcher or a research assistant in their classrooms. The observer went to the ITAs' classes two times over the course of a semester to observe the classes and completed the classroom observation rubric, which was used to evaluate each ITA's oral proficiency and teaching competence (see Appendix D). To maximize the validity and reliability of these teacher observation evaluations, it was decided to observe each ITA two times over the course of the semester, usually with about 1 month between observations.

Undergraduate Students' Evaluation of International Teaching Assistant Oral Proficiency

When the research observer went to the ITA's classroom to administer the classroom observation rubric, he or she also administered at the end of the class session the 22-item questionnaire (Appendix C) to the students in the class, asking the students about the instructor's oral proficiency and teaching competence. Again, this was administered to all the students in the class, two times during the semester.

Beginning-of-Semester and End-of-Semester TOEFL iBT Speaking and Listening Scores

Finally, a number of new ITAs were invited to participate in a specific part of the study focusing on Research Question 3, in which they took a modified (research) version of the TOEFL iBT at the beginning and at the end of their first semester as an ITA at Temple. ETS provided Form Creator software that allowed the researcher to administer retired versions of the TOEFL iBT for data collection purposes. A computer laboratory on campus was reserved, and participants were solicited to register for a test administration within the first 2 weeks of the opening of the semester and within 2 weeks of the close of the semester. Each of these test administrations took approximately 1 hour. The software program provided by ETS automatically scored the listening sections of the test. For the speaking component, the software created an audio (.wav) file of the test takers' spoken responses, and these files were submitted to ETS for rating. Two trained ETS raters scored each speaking sample.

ITAs who agreed to participate in the study and took both locally administered TOEFL iBT exams were compensated $100 for their time. Those ITAs who participated in the study by being observed and evaluated in the classroom two times were compensated $50 for their time.

Analyses

Research Question 1

To address Research Question 1 ("To what extent do TOEFL iBT Speaking and Listening scores correspond to other criterion measures of ITA teaching preparedness used at Temple University?"), first a correlation analysis was used. All of the potential ITAs submitted TOEFL iBT scores when applying for admission to Temple. Those incoming ITAs who scored less than 28 on the speaking section of the TOEFL iBT were also required to take the locally administered SPEAK test. The TOEFL iBT Speaking and Listening scores for all incoming ITAs who scored lower than 28 on the speaking section of the TOEFL iBT were correlated with their scores on the SPEAK test. Because the sets of scores are continuous variables, a Pearson product–moment correlation was used.

For those ITAs who scored lower than a 50 on the SPEAK test and thus were required to take ITA 5501 (and the culminating assessment, the TEACH test), a second analysis was conducted correlating the TOEFL iBT Speaking and TOEFL iBT Listening scores, SPEAK test scores, and TEACH test scores. Again, a Pearson product–moment correlation was used.

Table 4 Dependent and Independent Variables in the Regression Analyses

Regression	Type	*N*	Dependent (outcome) variable	Independent (predictor) variables
1	Standard	128	Teaching competence (SFF scores)	TOEFL iBT Listening scores, TOEFL iBT Speaking scores
2	Sequential	33	Teaching competence (student evaluation of ITA teaching competence)	TOEFL iBT Listening scores entered into regression equation after TOEFL iBT Speaking scores
3	Sequential	33	Teaching competence (observer evaluation of ITA teaching competence)	TOEFL iBT Listening scores entered into regression equation after TOEFL iBT Speaking scores

Note. ITA = international teaching assistant; SFF = student feedback form.

Research Question 2

To investigate Research Question 2 ("To what extent do the TOEFL iBT Speaking and Listening scores correspond to measures of ITAs' classroom teaching performance?"), a series of regression analyses was performed. The first analysis used existing data from all the ITAs who taught classes in their first semester as ITAs at Temple from 2006 to 2012. This included all ITAs who scored at least 28 on the TOEFL iBT and those who scored lower than 28 but received a score of 45 or higher ("pass" or "pass with restrictions") on the locally administered SPEAK test or those who scored less than 45 on the SPEAK test but enrolled in and successfully completed the ITA 5501 class. These participants' TOEFL iBT Speaking and Listening scores and SPEAK test scores were used as predictors of the outcome variable teaching competence, as operationalized by SFF scores.

All ITAs entering Temple in Fall 2009 or later were invited to participate in this study and to have their teaching observed and assessed by a rater who observed the ITA teaching in his or her own classroom and by the undergraduate students in the class (two times over the course of the semester). A total of 33 ITAs participated in this component of the study, and a second set of regression analyses was conducted in which TOEFL iBT Speaking and Listening scores were used as predictor variables of teaching competence, as operationalized by the outcome variables SFF evaluations and classroom observation scores. This information can be seen in Table 4.

Research Question 3

Finally, to address Research Question 3 ("What gains in oral proficiency [as measured by TOEFL iBT Speaking and Listening scores] do ITAs make after a semester of ITA training and/or teaching?"), it was necessary to have ITA participants take a research version of the TOEFL iBT Speaking and Listening sections. All new ITAs entering Temple in 2009–2012 who scored lower than 28 on the TOEFL iBT Speaking (and thus had to take the SPEAK test) were invited to take the Speaking and Listening sections of the TOEFL iBT within 2 weeks of their first semester as ITAs at Temple and also within 2 weeks of the end of their first semester as ITAs. The beginning-of-semester TOEFL iBT Speaking and Listening scores were compared to the end-of-semester TOEFL iBT Speaking and Listening scores to investigate the extent of the development in oral proficiency the ITAs demonstrated after a semester of living and studying in an English-speaking environment. A series of repeated measures analyses of variance (ANOVAs) was used to examine if the beginning-of-semester and end-of-semester TOEFL iBT Speaking and Listening scores differed significantly, and the effect sizes were calculated. The different analyses that were used to investigate each research question are shown in Table 5.

Results

Research Question 1

To investigate Research Question 1, a number of correlation analyses were conducted.

Correlation of TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK Test Scores

The TOEFL iBT Speaking and Listening scores were correlated with the SPEAK test scores for the 198 ITAs who had data for all three measures of interest: TOEFL iBT Speaking, TOEFL iBT Listening, and the SPEAK test. For these 198

Table 5 Data Analyses for Each Research Question

Research question	Data type	Analytical method	Variables
1	Existing	Correlation	TOEFL iBT Speaking scores; TOEFL iBT Listening scores; SPEAK scores; TEACH scores
2	Existing	Correlation	TOEFL iBT Speaking scores; TOEFL iBT Listening scores; SPEAK scores; TEACH scores; SFF evaluations (from first semester of teaching)
2	Existing	Regression	TOEFL iBT Speaking scores; TOEFL iBT Listening scores; SPEAK scores; SFF evaluations (from first semester of teaching)
2	Existing and newly collected	Regressions	TOEFL iBT Speaking scores; TOEFL iBT Listening scores; SPEAK scores; students' evaluation of ITA teaching competence; students' evaluation of ITA oral proficiency; observers' evaluation of ITA teaching competence; observers' evaluation of ITA oral proficiency
3	Newly collected	Repeated measures ANOVAs	TOEFL iBT Speaking scores (beginning of semester); TOEFL iBT Listening scores (beginning of semester); TOEFL iBT Speaking scores (end of semester); TOEFL iBT Listening scores (end of semester)

Table 6 TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK Test Scores for Overall Sample

Test	*N*	Maximum possible score	Minimum	Maximum	*M*	*SD*	Skewness	Kurtosis
TOEFL iBT Speaking	198	30	11	30	21.74	3.01	−.49	.06
TOEFL iBT Listening	198	30	9	30	24.04	3.88	.10	.69
SPEAK	198	60	35	60	47.58	5.69	−.05	−.36

Table 7 Correlations Between TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK Test Scores

	TOEFL iBT Speaking	TOEFL iBT Listening	SPEAK
TOEFL iBT Speaking	1		
TOEFL iBT Listening	.60**	1	
SPEAK	.71**	.59**	1

Note. $N = 198$.
**$p < .01$.

ITA participants, the mean score on the TOEFL iBT Speaking was 21.74 ($SD = 3.01$), with a minimum score of 11 and a maximum score of 30. The mean score on the TOEFL iBT Listening was higher (24.04, $SD = 3.88$), and there was more variation in the TOEFL iBT Listening scores, as noted by the higher standard deviation. These ITA participants had a mean score of 47.58 on the SPEAK test, with a standard deviation of 5.69, a minimum score of 35, and a maximum score of 60. This information is shown in Table 6.

Table 7 shows the correlations between TOEFL iBT Speaking scores, TOEFL iBT Listening scores, and the SPEAK test scores for the 198 ITAs. The correlation between TOEFL iBT Speaking and TOEFL iBT Listening was moderately high, at .60. The correlation between TOEFL iBT Speaking and the SPEAK test was also moderately high, at .71. The correlation between TOEFL iBT Listening and the SPEAK test was .59. All of these correlations are significant at the $p < .01$ level. The scatterplots for the correlations are provided in Appendix E.

Correlation of TOEFL iBT Speaking, TOEFL iBT Listening, the SPEAK Test, and TEACH Scores

From this sample of 198 ITAs, a subsample was used to investigate the correlations between TOEFL iBT Speaking, TOEFL iBT Listening, the SPEAK test, and TEACH test scores (TEACH Language, TEACH Teaching, and TEACH Composite). Again, 50 is the threshold score on the SPEAK test; a score of 50 or higher meant that the test taker was able to be an

Table 8 TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK Test, and TEACH Scores for Subsample

Test	*N*	Maximum possible score	Minimum	Maximum	*M*	*SD*	Skewness	Kurtosis
TOEFL iBT Speaking	89	30	11	24	20.02	2.52	−.49	.57
TOEFL iBT Listening	89	30	9	29	21.88	3.23	−.53	1.34
SPEAK	89	60	35	45	42.30	3.11	−.72	−.44
TEACH Language	89	20	10.95	20	15.63	2.15	−.07	−.56
TEACH Teaching	89	12	6.04	12	9.98	1.32	−.42	−.15
TEACH Composite	89	32	17	32	25.61	3.31	−.15	−.51

Table 9 Correlations Between TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK Test, and TEACH Scores

	TOEFL iBT Speaking	TOEFL iBT Listening	SPEAK	TEACH Teaching	TEACH Language	TEACH Composite
TOEFL iBT Speaking	1					
TOEFL iBT Listening	.56**	1				
SPEAK	.61**	.33**	1			
TEACH Teaching	.15	.34**	.30**	1		
TEACH Language	.40**	.39**	.39**	.81**	1	
TEACH Composite	.32**	.39**	.37**	.92**	.97**	1

Note. $N = 89$.

**$p < .01$.

instructor with no reservations and did not have to take the ITA class. Of the original sample of 198 ITAs described earlier, 90 scored 50 or higher and thus did not take the ITA class (and consequently did not have TEACH scores). Of the original 198 ITAs, 108 scored less than 50, but 19 chose not to take the ITA class (the main reason for this was that they were awarded a research assistantship rather than a teaching assistantship and thus did not need to take the ITA class). Thus, 89 ITAs remained of the original sample, so this subsample of 89 ITAs is given here. It must be noted that although the background demographics of this subsample of 89 ITA participants is similar to the background demographics of the larger sample of 198 ITA participants (as described in the "Participants" section), this subsample is different in that the participants have, by definition, lower speaking ability, as evidenced by their scores on the SPEAK test.

The mean score on the TOEFL iBT Speaking was 20.02 ($SD = 2.52$), with a minimum score of 11 and a maximum score of 24. The mean score on the TOEFL iBT Listening was higher at 21.88. There was also more variation in the TOEFL iBT Listening scores, with a minimum score of 9, a maximum score of 29, and a higher standard deviation of 3.23. These 89 ITAs had a mean score of 42.30 on the SPEAK test, with a standard deviation of 3.11, a minimum score of 35, and a maximum score of 45. The TEACH test assessed both language proficiency and teaching competence, so scores are given here for TEACH Language, TEACH Teaching, and TEACH Composite. The mean score on the TEACH Language was 15.63 (out of 20 possible), with a standard deviation of 2.15. The mean score on the TEACH Teaching was 9.98 (out of 12 possible), with a standard deviation of 1.32. The mean score on the TEACH Composite was 25.61 ($SD = 3.31$). This information is shown in Table 8.

The correlations between the TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK test, TEACH Language, TEACH Teaching, and TEACH Composite scores were computed and are shown in Table 9.

The correlation between TOEFL iBT Speaking scores and TEACH Composite scores was weak ($r = .32$, $p < .01$), as was the correlation between TOEFL iBT Speaking and TEACH Language scores ($r = .40$, $p < .01$). The correlation between TOEFL iBT Speaking and TEACH Teaching scores was very weak and not statistically significant. The correlations between TOEFL iBT Listening scores and the three sets of TEACH scores were all very similar: TOEFL iBT Listening correlated with TEACH Teaching ($r = .34$, $p < .01$); TOEFL iBT Listening with TEACH Language ($r = .39$, $p < .01$), and TOEFL iBT Listening with TEACH Composite ($r = .39$, $p < .01$). A sampling of the scatterplots for these correlations is given in Appendix E. In summary, five of the six correlations between the TOEFL iBT scores and the TEACH test scores were weak but statistically significant. The exception was between TOEFL iBT Speaking and TEACH Teaching, for which no correlation was found.

Discussion for Research Question 1

TOEFL iBT Speaking and TOEFL iBT Listening

For the entire group of 198 ITAs, the correlation between the TOEFL iBT Speaking and Listening scores was moderate ($r = .60$, $p < .01$). For the subset of 89 ITAs, the correlation was also moderately strong ($r = .56$, $p < .01$). The correlations between the TOEFL iBT Speaking and TOEFL iBT Listening scores were nearly the same for both sets of participants, even though the mean TOEFL iBT scores were higher for the entire group of 189 ITAs (TOEFL iBT Speaking, $M = 21.74$; TOEFL iBT Listening, $M = 24.04$) than the mean TOEFL iBT test scores for the 89-participant subset (TOEFL iBT Speaking, $M = 20.02$; TOEFL iBT Listening, $M = 21.88$).

TOEFL iBT Speaking and the SPEAK Test

For the overall sample of 198 ITAs, the TOEFL iBT Speaking scores correlated moderately with the SPEAK test scores ($r = .71$, $p < .01$). This correlation is marginally weaker than the correlations found in Xi's (2008) study, which found that the TOEFL iBT Speaking correlated strongly ($r = .78$) with the SPEAK test scores of the 84 ITA participants at UCLA and even more strongly ($r = .89$) with the SPEAK test scores of the 45 ITA participants at Drexel University. However, for the subsample of 89 ITA participants, the correlation of TOEFL iBT Speaking and SPEAK test scores was lower ($r = .61$, $p < .01$). Again, this subsample had a mean TOEFL iBT Speaking score approximately 1.71 points lower than the overall sample of 198 ITA participants, and the resulting slightly restricted range of scores and smaller standard deviation might have contributed to the somewhat lower correlation for this subsample.

Because the two tests are both semidirect tests of speaking ability that do not involve face-to-face interaction, it was expected that the two would correlate more strongly. However, as described earlier, there is concern about the security of the SPEAK test, and this lack of security could compromise the results of the test. In addition, because there is no ongoing training and norming of the SPEAK test raters, the reliability of the SPEAK test scoring is also in question. These factors might at least partly explain why there was not a stronger correlation between the TOEFL iBT Speaking and the SPEAK tests.

TOEFL iBT Listening and the SPEAK Test

For the larger sample of 198 ITAs, the TOEFL iBT Listening correlated moderately with the SPEAK test ($r = .59$, $p < .01$). The correlation between TOEFL iBT Listening and the SPEAK test scores for the subsample of 89 ITAs, however, was much lower ($r = .33$, $p < .01$). As with the correlation described previously between the TOEFL iBT Speaking and the SPEAK test, it is difficult to interpret this much lower correlation for the subsample of 89 ITAs. Because listening and speaking are both oral components of language, it was expected that they would correlate at least moderately strongly. Again, this subsample is by definition of lower proficiency, and they scored 2.16 points lower on the TOEFL iBT Listening than the overall sample of 198 ITAs; this restricted range of scores and lower standard deviation (as well as the security and scoring reliability issues with the SPEAK test) might at least partly explain the weak correlation.

TOEFL iBT Speaking and TEACH

For the subsample of 89 ITAs for which there were TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK test, and TEACH Language, Teaching, and Composite scores, TOEFL iBT Speaking correlated weakly with the TEACH Composite scores ($r = .32$, $p < .01$). However, when the TEACH scores were broken down into the language and teaching components, it was found that although the TOEFL iBT Speaking scores correlated weakly with the TEACH Language scores ($r = .40$, $p < .01$), TOEFL iBT Speaking scores did not correlate with TEACH Teaching scores ($r = .15$, n.s.).

It is difficult to interpret the weak correlation between TOEFL iBT Speaking scores and the TEACH Language scores, as a much higher correlation was expected. Indeed, this finding contrasts sharply with Xi's (2008) findings, who found a strong correlation between TOEFL iBT Speaking and the language component of the ITA screening exams at UCLA ($r = .78$), the University of North Carolina ($r = .73$), Drexel University ($r = .73$), and the University of Florida ($r = .48$). However, Oppenheim's (1998) study found an even lower correlation ($r = .33$) between the ITAs' scores on an oral English proficiency assessment and the linguistic subscale on a test of teaching effectiveness.

Although the rubrics for the TOEFL iBT Speaking and the TEACH Language tests are obviously different, a major component being assessed in both tests is presentational speaking ability. Again, the TEACH Language analytic rubric has five components (comprehensibility, accent and pronunciation, listening comprehension, fluency, and grammar and word choice). Although the listening comprehension subscale of the rubric obviously differs from the speaking abilities assessed in TOEFL iBT Speaking, the four other components of the analytic rubric all seem to be components of speaking also measured by TOEFL iBT Speaking.

However, the context of the TEACH test, in which the test takers taught a mini-lesson to a group of simulated students, is obviously very different from the context of the TOEFL iBT Speaking, which involves no human interaction. This teaching of a mini-lesson involves content knowledge, classroom cultural knowledge, pedagogical knowledge, the possibility of anxiety because of teaching, and a number of other factors that are not included in the TOEFL iBT Speaking, and it is difficult to separate speaking ability from the context and domain in which that speaking ability is assessed, as Schmidgall (2013) concluded. In addition, the teaching component of the TEACH test, though measured separately on the TEACH analytic rubric, still seems to be a factor in the language scores. For example, the test taker who paused to consider how to make the content of the lesson comprehensible to the simulated students might have been seen as less fluent (a scale on the rubric). Finally, the interactive nature of the TEACH test, in which the final 2 minutes of the mini-lesson were devoted to students asking questions of the teacher (although students were also able to ask questions at any time during the mini-lesson), allows for assessment of the test taker's interactive speaking and listening ability, which is not assessed by TOEFL iBT Speaking.

It is also difficult to interpret the lack of correlation between TOEFL iBT Speaking and the TEACH Teaching scores. Although it was expected that the correlation between TOEFL iBT Speaking and TEACH Teaching would be weaker than the correlation between TOEFL iBT Speaking and TEACH Language, it is surprising that the correlation ($r = .15$, n.s.) was so weak and not statistically significant. Xi (2008) found moderately strong correlations between TOEFL iBT Speaking and the teaching component of the ITA screening exams at the University of North Carolina ($r = .69$) and at Drexel University ($r = .52$). However, Xi also found only a very weak correlation on the University of Florida exam ($r = .14$). Oppenheim (1998) found a similarly low correlation ($r = .25$) between the ITAs' scores on an oral English proficiency assessment and a test of teaching effectiveness. Obviously, an instructor must have at least a certain level of oral proficiency in English to be a competent instructor in an English-medium university, but the finding here was no correlation between TOEFL iBT Speaking scores and TEACH Teaching scores. This issue is examined in more depth later, in relation to Research Question 3.

TOEFL iBT Listening and TEACH

Similar to the TOEFL iBT Speaking results, the TOEFL iBT Listening scores also correlated only weakly with the TEACH Composite scores ($r = .39$, $p < .01$). However, when the TEACH scores are divided into the language and teaching components, a different result occurs. As with the TOEFL iBT Speaking scores, the TOEFL iBT Listening scores correlated weakly with the TEACH Language scores ($r = .39$, $p < .01$), but the TOEFL iBT Listening scores also correlated weakly (and at a statistically significant level) with the TEACH Teaching scores ($r = .34$, $p < .01$). It is difficult to compare these results with other research, as there do not seem to be any studies in the literature examining the extent to which listening ability is correlated with or predicts teaching competence. Again, this is a relatively weak correlation, but it contrasts with the lack of correlation between TOEFL iBT Speaking and TEACH Teaching scores.

As noted earlier, the TEACH simulation involves interaction with students and a dedicated question-and-answer period. One of the five analytic components of the TEACH Language scale is listening comprehension; the correlation between the listening comprehension component of the TEACH Listening scale and the TOEFL iBT Listening is weak but statistically significant ($r = .29$, $p < .05$). However, it seems reasonable to believe that listening comprehension is also involved in a number of the analytic components of the TEACH Teaching rubric. Again, there are three analytic scales in the TEACH Teaching component of the rubric (lesson organization and implementation, relevance of content and development of content, and interaction with students and nonverbal communication). Certainly the scores the raters give for interaction with students is affected by listening comprehension, especially in relation to how the test taker responds to questions from the students. It seems likely that listening comprehension has some impact on the other analytic scales here as well.

Table 10 TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK Test, and Student Feedback Form (SFF) Scores

Test	*N*	Maximum possible score	Minimum	Maximum	*M*	*SD*	Skewness	Kurtosis
TOEFL iBT Speaking	128	30	15	30	22.06	3.16	.25	.10
TOEFL iBT Listening	128	30	16	30	24.58	3.71	−.34	−.74
SPEAK	128	60	35	60	47.70	5.98	.03	−.48
SFF scores[a]	128	5	2	4.76	3.86	.51	−.70	.66

[a] Average of three items.

These results provide some limited criterion-related evidence in support of the use of TOEFL iBT Listening scores for ITA screening purposes. Although the correlations between TOEFL iBT Listening and TEACH scores were relatively weak, they were statistically significant (as opposed to the lack of correlation between TOEFL iBT Speaking and TEACH test scores).

The Two Components of the TEACH Test

The two components of the TEACH test were highly correlated ($r = .81$, $p < .01$). This is identical to the correlation Xi (2008) found between the language and teaching components of the ITA screening exams at Drexel University ($r = .81$) and higher than the correlation she found at the University of North Carolina ($r = .61$); it contrasts dramatically with the very weak correlation between the teaching and language components on the University of Florida exam ($r = .10$). While the two components of the TEACH test were designed to assess two very different constructs (teaching competence and language competence), the way these two constructs were assessed might explain why they were so highly correlated. The raters of the TEACH test used a one-page rubric to give five different language scores and three different teaching scores. It seems highly likely that a halo effect occurred in this rating; that is, while giving eight separate scores for the same event, it seems likely the raters tended to give similar scores across the eight different categories. This was the conclusion of Wagner (2012), who examined the intercorrelations of the eight different analytic scales on the TEACH test and found that almost all were highly correlated (r ranged from a low of .58 to a high of .90, with 18 of the 21 correlations being larger than .70). Thus it seems likely that even though the TEACH analytic rubric had two different categories (teaching and language) and eight different analytic scales, this halo effect impacted how the assessors assigned scores for the different scales.

Research Question 2

To investigate Research Question 2, three regression analyses were performed on different subpopulations.

The first regression used existing data from all the ITAs from 2006 to 2012 from their first semester of teaching at Temple. This included all ITAs who scored at least 28 on the TOEFL iBT and those who scored lower than 28 but received a score of 45 or higher (pass or pass with restrictions) on the locally administered SPEAK test or those who scored below 45 on the SPEAK test but enrolled in and successfully completed the ITA 5501 class. There were 128 ITAs for which there were TOEFL iBT Speaking scores, TOEFL iBT Listening scores, SPEAK test scores, and SFF scores. The background characteristics of this subsample of 128 ITAs are similar to the characteristics of the overall 198-ITA sample (see the "Participants" section). The mean score on the TOEFL iBT Speaking was 22.06 ($SD = 3.16$), whereas the mean score on the TOEFL iBT Listening was higher at 24.58. These 128 ITAs had a mean score of 47.70 on the SPEAK test ($SD = 5.98$). Finally, for the SFF measures, students responded on a 5-point Likert-type scale to a number of questionnaire items. Three of these items were chosen to represent teaching competence, and the averages of these three items are given here on a 5-point scale, with 3 as the mid-point of the scale. Thus anything above 3 is considered as above average teaching, and anything below 3 is below average teaching. The mean SFF scores for these 128 participants was 3.86 ($SD = .51$). This information is shown in Table 10.

The correlations between the four variables (TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK, SFFs) are given in Table 11.

Although the correlations between the three predictor variables were all moderately strong ($r = .58$, $p < .01$ for TOEFL iBT Listening and TOEFL iBT Speaking; $r = .71$, $p < .01$ for the SPEAK test and TOEFL iBT Speaking; $r = .61$, $p < .01$ for the SPEAK test and TOEFL iBT Listening), the correlations between the dependent variable of SFF scores and the

Table 11 Correlations Between TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK Test, and Student Feedback Form (SFF) Scores

	TOEFL iBT Speaking	TOEFL iBT Listening	SPEAK	SFF scores
TOEFL iBT Speaking	1			
TOEFL iBT Listening	.58**	1		
SPEAK	.71**	.61**	1	
SFF scores	.12	.07	.06	1

Note. $N = 128$.
**$p < .01$.

Table 12 TOEFL iBT Speaking, TOEFL iBT Listening, Student Feedback Form (SFF) Scores, Observers' Assessment of Teaching Competence, and Students' Assessment of Teaching Competence Scores

Test	*N*	Maximum possible score	Minimum	Maximum	*M*	*SD*	Skewness	Kurtosis
TOEFL iBT Speaking	33	30	15	26	21.45	2.59	−.19	−.72
TOEFL iBT Listening	33	30	18	30	24.85	3.32	−.60	−.49
SFF scores[a]	33	5	1.83	4.73	3.79	.57	−1.38	1.56
Observers' assessment of teaching competence	33	4	1.75	3.88	2.91	.59	−.10	−.90
Students' assessment of teaching competence	33	5	2.29	4.38	3.85	.59	−2.06	1.55

[a]Average of three items.

explanatory variables were near zero and not statistically significant. Scatterplots for a sample of these correlations are provided in Appendix F.

A standard regression was then conducted to examine whether oral language proficiency (as measured by TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK test scores) would predict teaching competence, as measured by SFF scores. The results of the regression indicated an R^2 value of .017, meaning that the independent variables accounted for less than 2% of the variance in the SFF scores.

The regression analysis found that only a very small percentage of the ITAs' teaching competence (as measured by SFF scores) could be predicted by the ITAs' oral proficiency (as measured by TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK test scores). However, as is discussed in more detail later, using SFF scores as the sole measure of teaching competence is problematic for a number of reasons. Therefore, it was decided to include additional measures of teaching competence in this study. All ITAs entering Temple in Fall 2009 or later were invited to participate in this study and to have their teaching observed and assessed by a rater in the ITAs' own classrooms two times in a semester. In addition, the undergraduate students in the ITAs' classes completed a separate assessment (two times in a semester) of the ITAs' teaching competence that was more thorough and detailed than the SFFs. Thirty-three ITA participants had complete sets of data, including TOEFL iBT Speaking and Listening scores, students' assessments of the ITA's teaching competence, and observers' assessments of the ITA's teaching competence. These 33 ITAs were a subset of the 128 ITA participants used in the previous regression analysis, and the background characteristics of this subsample of 33 ITAs are similar to the characteristics of the 128-ITA sample. The mean score on the TOEFL iBT Speaking was 21.45 ($SD = 2.59$), while the mean score on the TOEFL iBT Listening was somewhat higher at 24.85 ($SD = 3.32$). The mean SFF scores for these 33 participants was 3.79 ($SD = .57$) (the average of the three Likert items on a 5-point scale). For the observers' assessment of teaching competence, which used the average across four subscales on a 4-point scale, the mean for these 33 participants was 2.91 ($SD = .59$). Finally, for the Students' assessment of teaching competence, students responded on a 5-point Likert-type scale to eight questionnaire items, with 3 as the mid-point of the scale. The mean score for these 33 participants was 3.85 ($SD = .59$). This information is shown in Table 12.

For these 33 participants, two additional regression analyses were conducted in which the measures of oral proficiency (as measured by TOEFL iBT Speaking and TOEFL iBT Listening scores) were used as predictors of teaching competence. For the first regression, the outcome variable was teaching competence as measured by student assessment of the ITA's teaching competence. For the second regression, the outcome variable was teaching competence as measured by observers' assessment of the ITA's teaching competence. Because TOEFL iBT Speaking is accepted as the standard predictor (i.e., a

Table 13 Correlations Between TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK Test, Student Feedback Form (SFF) Scores, and Observers' Assessment of ITAs' Language Proficiency and Teaching Competence

	TOEFL iBT Speaking	TOEFL iBT Listening	SPEAK	SFF scores	Observers' assessment of language proficiency	Observers' assessment of teaching competence	Students' assessment of language proficiency	Students' assessment of teaching competence
TOEFL iBT Speaking	1							
TOEFL iBT Listening	.59**	1						
SPEAK	.62**	.55**	1					
SFF scores	.01	.09	.04	1				
Observers' assessment of language proficiency	.35*	.58**	.48**	.22	1			
Observers' assessment of teaching competence	.04	.39*	.19	.26	.69**	1		
Students' assessment of language proficiency	.34	.62**	.26	.51**	.61**	.49**	1	
Students' assessment of teaching competence	.24	.46**	.29	.67**	.56**	.52**	.80**	1

Note. $N = 33$.
*$p < .05$. **$p < .01$.

high enough score on TOEFL iBT Speaking is widely accepted as indicating an oral proficiency necessary for teaching), a sequential regression was run to examine the effects of TOEFL iBT Listening after TOEFL iBT Speaking was added. The rationale for this is that TOEFL iBT Speaking is the standard measure used; this research question is trying to investigate if using TOEFL iBT Listening scores in addition to TOEFL iBT Speaking scores can result in better predictive power of teaching competence. Therefore these two regressions were sequential regressions, with TOEFL iBT Speaking scores entered first into the equation, followed by TOEFL iBT Listening scores.

Different sources provide different recommendations for the number of participants per explanatory variable in linear regression. Stevens (1996) recommended approximately 15 per variable, whereas Howell (2002) cited a rule of thumb of 10 observations for each variable. Because of the small number of ITA participants here ($n = 33$), it was necessary to include only the two predictor variables TOEFL iBT Speaking and TOEFL iBT Listening and the outcome variable (teaching competence). However, a number of different variables were examined to see the extent to which they correlated with each other for this subset of 33 ITA participants. These variables correlated were TOEFL iBT Speaking, TOEFL iBT Listening, SPEAK test scores, SFF scores, observers' assessment of ITAs' language proficiency, observers' assessment of ITAs' teaching competence, students' assessment of ITAs' language proficiency, and students' assessment of ITAs' teaching competence. These correlations can be seen in Table 13.

Although it is not feasible to examine all of the correlations in detail, a number of the statistically significant correlation coefficients are briefly examined here. First, the strongest correlation in the matrix is the correlation of students' assessment of ITAs' language proficiency and students' assessment of ITAs' teaching competence ($r = .80$, $p < .01$). Such a strong correlation suggests that the students were conflating language proficiency and teaching competence while they were filling out the questionnaire, further suggesting a halo effect in that they were rating these two constructs at the same time, on the same questionnaire. The second largest correlation coefficient found in the matrix is the correlation of observers' assessment of ITA language proficiency and observers' assessment of ITAs' teaching competence ($r = .69$, $p < .01$). Again, this suggests a halo effect, wherein the observer is rating the ITA's teaching competence and oral proficiency simultaneously. The correlation between the observers' assessment of ITAs' language proficiency and the students' assessment of ITAs' language proficiency was moderately strong ($r = .61$, $p < .01$).

The TOEFL iBT Speaking and TOEFL iBT Listening scores did not correlate with the SFFs ($r = .01$, n.s. for TOEFL iBT Speaking and SFFs; $r = .09$, n.s. for TOEFL iBT Listening and SFFs). In contrast, the correlations between the TOEFL iBT Speaking and TOEFL iBT Listening scores and the other measures of teaching competence (students' assessment of ITAs' teaching competence, observers' assessment of ITAs' teaching competence) are more varied. Whereas TOEFL iBT Speaking did not correlate with these measures (they were not statistically significant), TOEFL iBT Listening did correlate (weakly) with these two alternative measures of teaching competence ($r = .39$, $p < .05$ for TOEFL iBT Listening and

observers' assessment of ITAs' teaching competence; $r = .46$, $p < .01$ for TOEFL iBT Listening and students' assessment of ITAs' teaching competence). Finally, the correlation between the two teaching competence measures (observers' assessment of ITAs' teaching competence and students' assessment of ITAs' teaching competence) was moderate and statistically significant ($r = .52$, $p < .01$). In addition, the correlation between the two explanatory variables (TOEFL iBT Speaking and TOEFL iBT Listening) was also moderate and statistically significant ($r = .58$, $p < .01$). Scatterplots for a sample of these correlations are provided in Appendix G.

For the regression analyses, two sequential regressions were run to examine the effects of TOEFL iBT Listening when added after TOEFL iBT Speaking. For the first regression analysis, in which observers' assessment of ITAs' teaching was the teaching competence measure, TOEFL iBT Speaking accounted for very little explanatory power of the model ($R^2 = .009$). When TOEFL iBT Listening was added to the equation after TOEFL iBT Speaking, total R^2 for the model was .202, and the R^2 change was .201. In other words, TOEFL iBT Listening accounted for over 20% of the variance, even after TOEFL iBT Speaking was included.

There were similar results for the second regression analysis, in which students' assessment of ITAs' teaching was the teaching competence measure. The TOEFL iBT Speaking scores again accounted for little explanatory power of the model ($R^2 = .059$). When TOEFL iBT Listening was added to the equation after TOEFL iBT Speaking, however, the total R^2 for the model was .212, and the R^2 change was .152. In other words, TOEFL iBT Listening accounted for over 15% of the variance even after TOEFL iBT Speaking was included in the regression equation.

Discussion for Research Question 2

The results from these analyses clearly indicate that if the SFF data are used as the sole measure of teaching competence, the various measures of oral proficiency used in the study have almost no correlation with or predictive power of teaching competence. For the sample of 128 ITAs for whom there were SFF, TOEFL iBT Speaking, TOEFL iBT Listening, and the SPEAK test scores, the correlations among the oral proficiency measures and SFF data were all very weak, and none were statistically significant. The two oral proficiency measures (TOEFL iBT Speaking and TOEFL iBT Listening scores) could account for less than 2% of the variance in the SFF scores. Clearly these measures of oral proficiency are not good predictors of teaching competence, as measured by SFF scores. These results are similar to the results found by Bailey (1983) and Inglis (1993), who also investigated the correlation between ITAs' teaching competence (as measured by the equivalent of SFFs) and speaking proficiency (as measured by FSI Oral Interview scores by Bailey and the SPEAK test scores by Inglis). Bailey (1983) found no correlation between the ITAs' FSI scores and the students' evaluations of their teaching competence ($r = 18$, n.s.); Inglis (1993) reported almost identical results, with no correlation found between the ITAs' SPEAK test scores and their students' evaluations of their teaching competence ($r = .15$, n.s.).

However, using SFF data as the sole measure of teaching competence is problematic, for a variety of reasons. Although it is beyond the scope of this report to discuss in detail the limitations of the SFF data, the research is clear that undergraduate students' evaluations can be influenced by the students' personal beliefs and attitudes, interest in the class, perceived grade in the class, difficulty of the class, amount of work required by the class, and many other factors (Nelson, 1992; Rubin, 1992; Rubin & Smith, 1990; Yook & Albert, 1999). One obvious shortcoming of the SFFs is that they often serve as an evaluation of the class and the syllabus rather than of the individual instructor. Indeed, one of the (three) items used in the SFF data analyzed here was "I learned a great deal in this class." This can be perceived by the undergraduate students as an evaluation of the class itself and might have little to do with the individual instructor's teaching competence. The other two items in the SFFs that contributed to the SFF data were "The instructor taught this course well" and "The instructor was well organized and prepared for class." Obviously "The instructor taught this course well" is a direct evaluation of the instructor (and yet can still be influenced by some of the personal factors mentioned earlier). Finally, "The instructor was well organized and prepared for class" can be a direct evaluation of the instructor yet is still an imperfect measure of teaching competence, because a person can be well organized and prepared and still be a poor instructor (and vice versa). Again, end-of-course student evaluations of their instructors are obviously problematic.

For these reasons, it was decided in this study to include two other measures of teaching competence, two measures that were hoped to be more objective, thorough, reliable, and valid measures of teaching competence. With the students' assessment of ITAs' teaching competence, the students responded to eight items focusing on the teacher's skill in interacting with students, knowledge of American classroom culture, and ability to communicate content information. The perceived advantages to this assessment were that it included students' evaluation two times during the semester (rather

than one time at the end of the semester, as with the SFF), it involved eight items rather than just three (which should result in a more comprehensive and reliable assessment), and it focused solely on the instructor and not on the class itself. However, as with the SFF data, individual students could still have personal biases toward the instructor, toward the class, and so on, and thus it was also decided to include a more objective measure of teaching competence by using the observers' assessment of ITAs' teaching competence. Here a trained observer came into the classroom two times to observe and evaluate the instructor, using an analytic rubric with scores for lesson organization and preparedness, teacher–student interaction/classroom culture, delivery strategies/nonverbal communication, and communication of content information.

The results of the correlation and regression analyses with these two measures of teaching competence differed markedly from the results of the analyses using the SFF data as the measure of teaching competence. Similar to the data set involving the 128 ITAs in which SFF scores had essentially no correlation with the oral proficiency measures of TOEFL iBT Speaking, TOEFL iBT Listening, and SPEAK test scores, the data with these 33 ITA participants essentially found no correlation between these measures either. In addition, the correlations between the spoken language proficiency measures (TOEFL iBT Speaking and SPEAK test scores) and these alternative teaching competence measures (observers' assessment of teaching competence and students' assessment of teaching competence) were very weak and not statistically significant.

In contrast, the TOEFL iBT Listening scores and these two measures of teaching competence had statistically significant correlations. This correlation of TOEFL iBT Listening scores with the two measures of teaching competence was confirmed with the two regression analyses. While the TOEFL iBT Speaking scores accounted for less than 1% of the variance of the observers' assessment of teaching competence, the TOEFL iBT Listening scores accounted for an additional 20% of the variance. Similarly, while the TOEFL iBT Speaking scores accounted for only 5.9% of the variance for the students' assessment of teaching competence, the TOEFL iBT Listening scores accounted for an additional 15.3% of the variance. Again, it is necessary to point out that TOEFL iBT Listening was added to the regression equation after TOEFL iBT Speaking, because TOEFL iBT Speaking is typically the measure used for ITA oral proficiency. The point of this part of the study was to examine if using TOEFL iBT Listening in addition to TOEFL iBT Speaking would provide useful predictive information. In fact, these results indicate that for this population, TOEFL iBT Speaking does not predict teaching competence, whereas TOEFL iBT Listening does. These results are even more striking when considering that speaking and listening ability are obviously related, and the correlation between TOEFL iBT Listening and TOEFL iBT Speaking for these 33 participants was moderately strong ($r = .58$, $p < .01$). Even with this moderately strong correlation between the two predictor variables, the TOEFL iBT Listening scores added unique variance to the prediction of teaching competence. These results suggest that listening ability does play a part in teaching competence and in turn that TOEFL iBT Listening scores can and should be used as a measure of oral language proficiency for ITA screening purposes.

This idea of using TOEFL iBT Listening scores in addition to TOEFL iBT Speaking scores for ITA screening purposes is sound theoretically. Obviously, to be an effective teacher in an English-medium university, the instructor has to have a certain level of oral proficiency in English. It also seems obvious that the oral proficiency needed to be an effective instructor involves more than just speaking ability. Except perhaps for a small percentage of classes that are strictly lecture, virtually all instructors interact orally with students and respond to student questions, and numerous ITA studies have concluded that these aspects are part of the ITA instructional TLU domain (Gorsuch, 2003, 2006; Hoekje & Linnell, 1994; Papajohn, 2006; Saif, 2002). Plough et al. (2010) found listening comprehension as the single biggest predictor of success on the Graduate Student Instructors Oral English Test, which is used for ITA screening. Elder (1993) found that IELTS Listening scores correlated more highly with success in a teacher education program than did any other components of IELTS scores (global, reading, writing, speaking). Myers (1994) specifically described how important students' questions are in many instructional contexts that ITAs are often in (i.e., labs). Similarly, Plakans (1997) found that one of the most common complaints of undergraduate students was when the ITA instructor was not able to understand and reply appropriately to student questions. She described how the language undergraduate students used in interacting with their ITA instructors is informal and different from the type of language ITAs were exposed to in their EFL classes when they learned English in their home countries. Plakans concluded that oral proficiency tests used in ITA screening "may need to include more opportunities for the examinees to demonstrate their listening skills and ability to respond to typically informal student talk" (p. 110). Part of the ability to respond appropriately to questions and "informal student talk" is related to the listener's ability to use the nonverbal components of interactive spoken language (Wagner, 2010a, 2013, 2014a). Indeed, this is a skill focused on in ITA 5501 and is a component of the rubric for the TEACH test and part of the rubric for observers' assessment of teaching competence used here.

Using TOEFL iBT Listening scores in addition to TOEFL iBT Speaking scores for ITA screening purposes also is logical from a practical standpoint. Powers and Powers (2015) found that using all four parts of the *TOEIC*® test (reading, writing, listening, and speaking) resulted in better predictions of test takers' "real-life" English skills in a particular domain (e.g., writing) than using only the section score (e.g., TOEIC Writing) for that domain. They argued that although the language domains are distinct, they are still strongly correlated, and that more information is almost invariably better than less information in predicting competence. They thus concluded that "a more precise estimate of English proficiency in a specific language domain is possible by assessing skills not only in that domain, but in other related domains as well" (p. 161). They qualified this claim by noting that the increased precision is due mainly to the first additional measure that was added to the regression and that diminishing returns were found after this first additional measure was added. Again, for an ITA teaching TLU domain, using both TOEFL iBT Speaking and TOEFL iBT Listening scores should result in a better prediction of teaching competence. However, it would not seem to be appropriate to include TOEFL iBT Reading and TOEFL iBT Writing scores, because they seem much less relevant for a teaching TLU domain, and their inclusion might in fact mask an ITA's possible lack of oral proficiency.

In addition, it needs to be acknowledged that logically, measures of oral proficiency in English might not be expected to be strong predictors of teaching competence (as measured by outcome variables such as SFFs, students' assessment of ITAs' teaching competence, and observers' assessment of ITAs' teaching competence). To be a competent instructor at an English-medium university, at least some threshold of English proficiency is required. Yet the SFFs are designed to measure teaching competence, and though language proficiency is a necessary component of teaching competence, it is only a part. Indeed, Cho and Bridgeman (2012) described how a number of studies have found that TOEFL iBT scores by themselves are not a strong predictor of subsequent academic success. In their study, Cho and Bridgeman found that the TOEFL iBT could explain only about 4% of variance in the GPAs of the graduate students in their study and only about 3% of the variance in the GPAs of the undergraduate students in their study. They concluded that "the nature of the relationship between language proficiency and academic success is complex and difficult to demonstrate" (p. 422). It should also be noted, however, that Cho and Bridgeman (2012) utilized additional, noncorrelational analyses to unpack the complex nature of the relationship, and reported that even small correlations " … might indicate a meaningful relationship between TOEFL iBT scores and GPA" (p. 421).

Similarly, Graham (1987) reviewed a number of studies that used language ability (usually operationalized as language proficiency test scores) as a predictor of academic success (generally operationalized as GPA) and found nonsignificant correlations between language proficiency and academic success for roughly half the studies reviewed, and for the other half of the studies, the correlations between the two measures were only weak to moderate. Criper and Davies (as cited in Elder, 1993) concluded that .30 is probably as high a correlation as can be expected for a language test as a predictor of academic success, because there are so many nonlinguistic factors that dictate academic success. It would seem that teaching success, in comparison to academic success, is even a step further removed from English proficiency. Elder (1993) described how there are so many nonlanguage variables (e.g., subject knowledge, interpersonal skills, cultural competence) that are involved in teaching competence, and thus it might not be considered surprising that the TOEFL iBT test scores in this study did not correlate with or predict teaching competence as measured by SFF scores. Yet the TOEFL iBT Listening scores did correlate with and predict the other two measures of teaching competence used in this study (observers' assessment of ITAs' teaching competence and students' assessment of ITAs' teaching competence), providing empirical evidence in support of the use of TOEFL iBT Listening scores in ITA placement decisions.

Research Question 3

To investigate Research Question 3, a series of repeated measures ANOVAs was conducted to examine if there was a statistically significant difference between beginning-of-semester TOEFL iBT Speaking scores and end-of-semester TOEFL iBT Speaking scores and between beginning-of-semester TOEFL iBT Listening scores and end-of-semester TOEFL iBT Listening scores.

A total of 84 ITAs took the TOEFL iBT at the beginning and end of the first semester of ITA activities. Of these 84 ITAs, all were in their first semester of ITA activities. However, "first semester of ITA activities" could mean a number of different things. For example, of the 84 ITAs, 55 were teaching a class, whereas 29 were not. Of the 84 ITAs, 40 were taking the ITA 5501 class. Of these 40, 11 were also teaching a class (indicating that they scored 45 on the SPEAK test with provisional passes), whereas 29 were not teaching a class that first semester (18 scored 40 or lower on the SPEAK test,

Table 14 Gender, Age, and Length in United States for Different International Teaching Assistant Subgroups

Group	*N*	Men	Women	Mean age (years)	Years in United States (mean/median)
Entire group	84	50	34	26.21	.83/.50
Taking ITA 5501	40	25	15	26.18	.83/.50
Not taking ITA 5501	44	25	19	26.25	.84/.50
New to United States	41	23	18	24.90	.11/.10
Long term in United States	43	27	16	27.46	1.52/1.20

Table 15 Beginning-of-Semester and End-of-Semester TOEFL iBT Speaking and TOEFL iBT Listening Scores for Group of 84 International Teaching Assistants (ITAs)

Group	*N*	Beginning-of-semester mean score[a] (*SD*)	End-of-semester mean score[a] (*SD*)	Mean difference from beginning to end of semester[b]	*F* value (*df*)	Sig.	Effect size *D*
All ITAs TOEFL iBT Listening	84	24.40 (4.33)	25.05 (3.99)	.65	3.31 (1, 83)	$p = .073$	.155
All ITAs TOEFL iBT Speaking	84	21.44 (3.68)	22.43 (3.46)	.99	22.03 (1, 83)	$p < .001$	.277

[a]Out of 30. [b]On a 30-point scale.

whereas 11 scored 45 but were not teaching). In addition, while this was the first semester of ITA activities for all of the participants, this does not necessarily mean that they were all new to the United States or even to Temple. For example, 41 reported being new to the United States (having lived there for 2 months or less), whereas 43 reported having lived in the United States for at least 5 months. Some of these 43 had studied at Temple, but this was their first semester as an ITA or ITA-in-training, whereas others had lived in the United States prior to studying at Temple. Of the 41 ITAs in the "new to the United States" group, 19 were taking the ITA 5501 class and 22 were not, and of the 43 ITAs in the "long term in the United States" group, 21 were taking the ITA 5501 class and 22 were not. The demographic data about these 84 participants are provided in Table 14 and include information about gender, age, and years in the United States at the time of participation in this study (when the ITAs took the beginning-of-semester TOEFL iBT exam).

These demographic data indicate that the various subgroups are similar according to the percentages of women versus men (there are more men in each of the groups). The average ages for the groupings of "taking ITA 5501" and "not taking ITA 5501" are almost identical. The one notable difference in group demographic characteristics is age: The average age for the new to the United States subgroup is approximately 2.5 years less than for the long term in the United States group.

A series of repeated measures ANOVAs was conducted to examine if there were differences in beginning-of-semester and end-of-semester scores for the various groups: (a) entire group ($N = 84$) beginning-of-semester TOEFL iBT Listening scores compared to end-of-semester TOEFL iBT Listening scores; (b) entire group ($N = 84$) beginning-of-semester TOEFL iBT Speaking scores compared to end-of-semester TOEFL iBT Speaking scores; (c) taking ITA 5501 class ($n = 40$) beginning-of-semester TOEFL iBT Listening scores compared to end-of-semester TOEFL iBT Listening scores; (d) not taking ITA 5501 class ($n = 44$) beginning-of-semester TOEFL iBT Listening scores compared to end-of-semester TOEFL iBT Listening scores; (e) new to the United States ($n = 41$) beginning-of-semester TOEFL iBT Speaking scores compared to end-of-semester TOEFL iBT Speaking scores; and (f) long term in the United States ($n = 43$) beginning-of-semester TOEFL iBT Listening scores compared to end-of-semester TOEFL iBT Listening scores.

Overall Group of 84 International Teaching Assistants

The difference in the TOEFL iBT Listening scores and TOEFL iBT Speaking scores from Time 1 (beginning of semester) and Time 2 (end of semester) for the entire sample of 84 ITAs was tested with a repeated measures ANOVA. Results are shown in Table 15.

The results show a nonstatistically significant effect for the difference between the beginning-of-semester and end-of-semester TOEFL iBT Listening scores. The group of 84 ITAs scored slightly higher (0.65 points on a 30-point scale) on

Table 16 Beginning-of-Semester and End-of-Semester TOEFL iBT Speaking and TOEFL iBT Listening Scores for International Teaching Assistants (ITAs) Taking or Not Taking ITA 5501

Group	*N*	Beginning-of-semester mean score[a] (*SD*)	End-of-semester mean score[a] (*SD*)	Mean difference from beginning to end of semester[b]
ITAs taking ITA 5501, TOEFL iBT Listening	40	22.28 (4.52)	23.25 (4.25)	.97
ITAs not taking ITA 5501, TOEFL iBT Listening	44	26.34 (3.10)	26.68 (2.94)	.34
ITAs taking ITA 5501, TOEFL iBT Speaking	40	19.28 (2.74)	20.25 (2.84)	.97
ITAs not taking ITA 5501, TOEFL iBT Speaking	44	23.41 (3.32)	24.41 (2.71)	1.00

[a]Out of 30. [b]On a 30-point scale.

the end-of-semester TOEFL iBT Listening test than they did on the beginning-of-semester TOEFL iBT Listening test, although this difference in scores was not statistically significant ($F_{1,83} = 3.31$, $p = .073$). In contrast, the results show a small effect size and statistically significant effect for the difference between the beginning-of-semester and end-of-semester TOEFL iBT Speaking scores. The group mean on the end-of-semester TOEFL iBT Speaking test was 0.99 points higher than the group mean on the beginning-of-semester TOEFL iBT Speaking test, and this difference was statistically significant ($F_{1,83} = 22.03$, $p < .001$).

International Teaching Assistants Taking ITA 5501 Class

Of the overall group of 84 ITAs who took the beginning-of-semester and end-of-semester TOEFL iBT Speaking and TOEFL iBT Listening tests, 40 were enrolled in the ITA preparation class (ITA 5501) and 44 were not enrolled in the class. To investigate whether these two groups differed in their growth in speaking and listening proficiency, another series of repeated measures ANOVAs was conducted. The mean scores on both TOEFL iBT Listening and TOEFL iBT Speaking are shown in Table 16.

The group of 40 ITAs taking ITA 5501 scored almost a full point higher on the end-of-semester TOEFL iBT Listening test than they did on the beginning-of-semester TOEFL iBT Listening test. For the group of 44 ITAs not taking ITA 5501, however, the difference between the end-of-semester and beginning-of-semester TOEFL iBT Listening scores was smaller, just .34 points. To examine if these score differences were statistically significant, a repeated measures ANOVA was conducted. The results show that there was not an overall statistical main effect for TOEFL iBT Listening score gains ($F_{1,82} = 3.30$, $p = .073$). In addition, there was no interaction effect ($F = .801$, $p = .374$) between the score gains and group (taking ITA 5501/not taking ITA 5501), suggesting that neither of the groups scored significantly higher on the end-of-semester TOEFL iBT Listening test than they did on the beginning-of-semester TOEFL iBT Listening test. The effect sizes were computed, and there was a small effect size for both groups ($D = .244$ for ITAs taking 5501; $D = .113$ for ITAs not taking ITA 5501), but again, the differences in scores were not statistically significant.

In contrast, the results of the repeated measures ANOVA for TOEFL iBT Speaking scores indicated that there was an overall statistical main effect for TOEFL iBT Speaking score gains ($F_{1,82} = 21.77$, $p < .001$). The group of ITAs taking ITA 5501 scored .97 points higher on the end-of-semester TOEFL iBT Speaking test, and the group of ITAs not taking ITA 5501 scored 1.00 points higher. There was no interaction effect ($F_{1,82} = .003$, $p = .953$) between the score gains and group (taking ITA 5501/not taking ITA 5501), suggesting that both groups scored significantly higher on the end-of-semester TOEFL iBT Speaking test than they did on the beginning-of-semester TOEFL iBT Speaking test. The effect sizes were also computed, and there was a small effect size for both groups ($D = .348$ for ITAs taking ITA 5501; $D = .330$ for ITAs not taking ITA 5501). It is worth noting that the score gains for the group of ITAs taking ITA 5501 were identical (.97 points) for TOEFL iBT Listening and TOEFL iBT Speaking, yet the difference was only statistically significant for TOEFL iBT Speaking, because the standard deviations for TOEFL iBT Speaking were much lower than they were for TOEFL iBT Listening.

International Teaching Assistants New to the United States

Of the overall group of 84 ITAs who took the beginning-of-semester and end-of-semester TOEFL iBT Speaking and Listening tests, 41 were new to the United States (had lived there less than 2 months), and 43 were longer term in the

Table 17 Beginning-of-Semester and End-of-Semester TOEFL iBT Speaking and TOEFL iBT Listening Scores for International Teaching Assistants (ITAs) New to or Long Term in the United States

Group	*N*	Beginning-of-semester mean score[a] (*SD*)	End-of-semester mean score[a] (*SD*)	Difference in group mean from beginning to end of semester[b]
ITAs new to United States, TOEFL iBT Listening	41	24.34 (4.28)	24.98 (4.01)	.64
ITAs long term in United States, TOEFL iBT Listening	43	24.47 (4.43)	25.12 (4.02)	.75
ITAs new to United States, TOEFL iBT Speaking	41	21.17 (3.60)	22.46 (3.36)	1.29
ITAs long term in United States, TOEFL iBT Speaking	43	21.70 (3.78)	22.40 (3.59)	.70

[a]Out of 30. [b]On a 30-point scale.

United States (had lived there at least 4 months). To investigate whether these two groups differed in their development in speaking and listening proficiency, another series of repeated measures ANOVA was conducted. The mean scores on both TOEFL iBT Listening and TOEFL iBT Speaking are shown in Table 17.

The 41 ITAs new to the United States scored .64 points higher on the end-of-semester TOEFL iBT Listening test than on the beginning-of-semester TOEFL iBT Listening test. Similarly, the 43 ITAs long term in the United States scored .75 points higher on the end-of-semester TOEFL iBT Listening test than on the beginning-of-semester TOEFL iBT Listening test. To examine if these score differences were statistically significant, a repeated measures ANOVA was conducted. The results show that there was not an overall statistical main effect for TOEFL iBT Listening score gains ($F_{1,82} = 3.267$, $p = .074$). In addition, there was no interaction effect ($F = .001$, $p = .981$) between the score gains and group (new to the United States/long term in the United States), suggesting that neither of the groups scored significantly higher on the end-of-semester TOEFL iBT Listening test than they did on the beginning-of-semester TOEFL iBT Listening test. The effect sizes were computed, and the effect sizes for both groups were very small ($D = .154$ for ITAs new to the United States; $D = .155$ for ITAs long term in the United States), but again, the difference in scores were not statistically significant.

In contrast, the results of the repeated measures ANOVA for TOEFL iBT Speaking scores indicated that there was an overall statistical main effect for TOEFL iBT Speaking score gains ($F_{1,82} = 21.765$, $p < .001$). The group of ITAs new to the United States scored 1.29 points higher on the end-of-semester TOEFL iBT Speaking test, and the group of ITAs not taking ITA 5501 scored .70 points higher. There was no interaction effect ($F = 2.021$, $p = .159$) between the score gains and group (new to the United States/long term in the United States), suggesting that both groups scored significantly higher on the end-of-semester TOEFL iBT Speaking test than they did on the beginning-of-semester TOEFL iBT Speaking test. The effect sizes were also computed, and there was a small effect size for the new to the United States group ($D = .370$) and a small effect size for the long term in the United States group ($D = .190$).

Discussion for Research Question 3

The results of the series of repeated measures ANOVAs are relatively straightforward; the 84 ITA participants in this sample improved in their speaking ability over the course of their first semester of ITA activity as measured by the speaking section of the TOEFL iBT. In contrast, the 84 ITAs did not improve in their listening ability, as measured by TOEFL iBT Listening. When broken down into subgroups, the results are somewhat more varied. The 40 ITAs who were taking ITA 5501 scored .97 points higher on the end-of-semester TOEFL iBT Listening test, although this was not a statistically significant difference. Those 44 ITAs not taking ITA 5501 scored only .34 points higher on the end-of-semester TOEFL iBT Listening test, and the difference was not statistically significant. The ITAs taking ITA 5501 scored .97 points higher on the end-of-semester TOEFL iBT Speaking test, and those ITAs not taking ITA 5501 also scored higher (1.00 points) on the end-of-semester TOEFL iBT Speaking test; both of these differences were statistically significant. Finally, when examining how length of residence in the United States might affect oral proficiency development, neither the ITAs new to the United States nor those long term in the United States scored significantly higher on the end-of-semester TOEFL iBT Listening test. In contrast, both subgroups scored significantly higher on the end-of-semester TOEFL iBT Speaking test. The 41 ITAs new to the United States scored 1.29 points higher, whereas those ITAs long term in the United States scored .70 points higher. Although both of these differences were statistically significant, both the magnitude of the gain

score and the effect size ($D = .370$) were larger for the group of ITAs new to the United States than for those who had been in the United States long term ($D = .190$).

Before examining possible reasons for these findings, it must be noted that although the gains in the TOEFL iBT Speaking scores are statistically significant, they are still relatively small — sometimes less than 1 point on a 30-point scale. It must also be acknowledged that these gains might be affected by a repeater effect: The participants took the two versions of the TOEFL iBT on average about 3 months apart. Although there is a common belief that test takers tend to improve their scores as they become more familiar with the format of the test tasks, there is little empirical evidence of this in relation to language proficiency tests (Kokhan & Lin, 2014). Indeed, Zhang (2008) investigated the reliability and validity of the TOEFL iBT by performing repeater analyses on the scores of 12,385 test takers who took the TOEFL iBT two times within a 1-month period. Zhang found that the increase in scores from the first test to the second test was negligible for the overall test and that the changes in the scores on the speaking and listening components of the TOEFL iBT were smaller than the changes on the reading and writing components of the test. Again, the average duration between tests for the participants in this study was 3 months, and so the time between the tests would seem to minimize a repeater effect. In addition, the participants in this study took different versions of the TOEFL iBT to ensure that they were not familiar with the specific content of the test. Finally, all of the participants in this study had already taken the TOEFL iBT prior to taking the research versions for this study and thus were already familiar with the TOEFL iBT tasks. For these reasons, the increase in scores due to test familiarity or a repeater effect appear to be minimal.

The fact that the score gains for TOEFL iBT Listening were not statistically significant and the score gains for TOEFL iBT Speaking after a semester of study were small might seem surprising, and even disappointing. These ITAs are immersed in an English-speaking environment. They are taking classes in English, studying in English, and interacting in English with students, faculty, staff, and other people on a daily basis. In addition, half of the ITAs in this sample were also taking ITA 5501, which is specifically designed to improve their speaking and listening ability in English. Nevertheless, their gains in speaking ability, as measured by the TOEFL iBT Speaking scores, are relatively small (only about 1 point on the 30-point TOEFL iBT scale), and their gains in listening ability were even smaller.

Yet these findings are very much in line with much of the literature. Numerous studies (e.g., Derwing et al., 2008; Freed, 1995; Segalowitz & Freed, 2004; Towell et al., 1996) have shown that development in oral proficiency is surprisingly limited, even in an immersion setting. Elder and O'Loughlin (2003) investigated the band score gains on the IELTS after a 10- to 12-week intensive English-language study program in Australia and New Zealand. The overall findings were that the students made very limited gains in their English proficiency. In contrast to the current study, Elder and O'Loughlin found that listening was the skill with the greatest amount of gain. Similarly, Ling et al. (2014) found that the 111 participants in their study showed much larger gains in TOEFL iBT Listening scores (5.63 points) than gains in TOEFL iBT Speaking scores (1.97 points).

Again, although these small gains can be seen as disappointingly small, the literature has suggested that this is often the case, and the literature provides a number of reasons for why this is so. Learning a second language is a long process, and the 3-month period between tests for this study is very short (Brecht, Davidson, & Ginsberg, 1995; Elder & O'Loughlin, 2003; Freed 1995; Huebner, 1995; O'Brien, Segalowitz, Freed, & Collentine, 2007; Segalowitz & Freed, 2004). In addition, these ITAs had a relatively high level of proficiency already (24.4 on the TOEFL iBT Listening, 21.4 on the TOEFL iBT Speaking), and the literature has consistently shown that dramatic gains in language proficiency are less likely in more advanced L2 learners than in beginning L2 learners (Brecht et al., 1995; Brecht & Robinson, 1995; Freed, 1995; Lapkin, Hart, & Swain, 1995; Llanes & Munoz, 2009). Ling et al. (2014) found much larger gains in English proficiency in their study, but it is difficult to compare their results with the results found here, because their study involved a longer period of learning (6 months for one of the groups and 9 months for the other group) and also lower-level learners (their group of 111 learners improved from 8.17 to 13.80 on the TOEFL iBT Listening and from 13.32 to 15.25 on the TOEFL iBT Speaking).

Another possible reason for the seemingly small gains is that international students on North American university campuses often have little interaction in English outside of class. Ranta and Meckelborg (2013) found that many of the international graduate students in their study (including many who were ITAs) had very limited interaction with native English speakers, even though they were taking (and sometimes teaching) classes and living in an English-speaking environment. Numerous studies (e.g., Cheng & Fox, 2008; Collentine & Freed, 2004; Freed, 1995; Lapkin et al., 1995; Myles & Cheng, 2003; Regan, 1995; Segalowitz & Freed, 2004; Trice, 2004) have found similar results: that L2 learners living in

target-language environments have less exposure to the target language than might be expected. Finally, it should be noted that, similar to O'Loughlin and Arkoudis (2009), there was a great deal of individual variation in the scores. Some ITAs made large gains, whereas others actually scored lower on the end-of-semester tests. Taking these factors into account, it could be argued that the fact that these ITAs had a measurable increase in their speaking proficiency in such a short time period is actually encouraging. Mastering a second language can be a long and arduous process, and the fact that these ITAs made measurable gains in their TOEFL iBT Speaking scores after only a 3-month period can be seen as definite progress.

However, the fact that these gains in oral proficiency were so limited over the course of the semester has implications for ITA programs. The whole notion of cut scores for ITA programs is that those potential ITAs who do not score high enough on the initial screenings (i.e., TOEFL iBT Speaking) are deemed as not having the necessary oral proficiency in English to be an effective instructor. Therefore these potential ITAs must take an ITA preparation course that will focus in part on improving the ITAs' oral proficiency so that they will develop the language skills necessary to be effective instructors. Indeed, the group of ITAs in this study enrolled in ITA 5501 did improve in their TOEFL iBT Speaking scores. The question remains, however, if their oral proficiency has improved enough for them to be effective instructors. If we look solely at their TOEFL iBT test scores, the answer to this question is probably no. The 40 ITAs enrolled in ITA 5501 scored a mean of 23.35 on the TOEFL iBT Listening at the end of their ITA 5501 semester and 20.25 on the TOEFL iBT Speaking. Again, though there was improvement, these scores gains were relatively small, and the end-of-semester scores are still relatively low, certainly much lower than the designated cut score (28 on TOEFL iBT Speaking), and lower than the requirements at almost all North American universities.

But what these ITAs learned in ITA 5501 included many language skills that are not measured by the TOEFL iBT, including how to respond appropriately to questions, how to engage in interactive speaking and listening, how to listen to and comprehend unplanned spoken discourse, and how to utilize the nonverbal communication of speakers. A number of ITA studies (e.g., Myers, 1994; Plakans, 1997) have found that being able to respond appropriately to students' questions is one of the most valued language skills students see in their instructors. Interactive, two-way conversation is one of the most typical types of real-world language contexts, yet this is the type of language task that English learners often find most difficult and for which they are often least prepared, because their formal instruction has focused on one-way presentational speaking and one-way interpretive listening (Wagner, 2014a). In addition, although TOEFL iBT Speaking includes tasks that involve responding to written and spoken input, truly interactive oral language use is not tested by the TOEFL iBT (Farnsworth & Wagner, 2013; Wagner, 2012, 2014a). Yet in many classroom contexts, including laboratory classes, where many of the ITAs eventually teach, truly interactive, two-way conversations between instructors and students are common, and the instructor's ability to engage in these interactive conversations is vital.

Related to this notion of interactive, two-way conversations is the type of language spoken. Real-world, unplanned spoken discourse is often very different from the types of spoken texts that are used in L2 classrooms and assessments. Authentic, unplanned spoken discourse tends to have connected speech and other phonological modifications that are often lacking in spoken texts specially prepared for L2 learners and test takers (Wagner, 2014a, 2014b). Authentic, unplanned spoken discourse also differs from these scripted and specially prepared spoken texts in that it tends to have more hesitation phenomena (filled and unfilled pauses, false starts, redundancies), have more slang and colloquial language, be less logically organized, and follow spoken grammatical norms (Gilmore, 2007; Wagner, 2014a, 2014b). Wagner and Toth (2014) found that L2 learners vary in their ability to comprehend this authentic spoken discourse. Yet the TOEFL iBT Listening test uses spoken texts that are entirely scripted and that do not have many of these characteristics of unplanned spoken language (Wagner, 2016); thus the test might not capture the test takers' varying levels of ability to process and comprehend authentic, unplanned spoken language.

In contrast, the language focus of the ITA 5501 class includes lessons devoted to understanding authentic, unplanned spoken discourse. Indeed, the textbook used in the class is *English Communication for International Teaching Assistants* (Gorsuch, Myers, Pickering, & Griffee, 2013), which specifically focuses on discourse intonation and the use of authentic speech samples to improve the ITAs' communicative speaking and listening abilities, which again might not be fully captured with the TOEFL iBT. Similarly, Wagner (2008, 2010a, 2013b) has found that L2 learners vary in their ability to utilize the nonverbal components of spoken language. Some L2 learners are better able than others to use this information to comprehend spoken input, but again, the TOEFL iBT does not capture this variance. As stated, the curriculum for ITA 5501 includes a focus on these components of communicative language ability. In addition, living in an English-speaking

environment and interacting with other speakers of English in real-world contexts (as compared to the English learning situations the ITAs experienced in their EFL classrooms in their home countries) would seem to lead to improvement in these aspects of communicative language ability, even though this improvement might not be measured by the TOEFL iBT tasks.

To return to the original question about whether the oral proficiency of those prospective ITAs has improved enough for them to be effective instructors, the small gains found in their TOEFL iBT Speaking and Listening scores might not be indicative of their improvement in overall communicative language ability. This is especially relevant for the TOEFL iBT Listening test. Whereas the TOEFL iBT Speaking test seems to have more validity evidence supporting its ability to capture and measure test takers' communicative speaking ability (i.e., Biber & Gray, 2013; Bridgeman et al., 2012), there seems to be less validity evidence supporting the ability of the TOEFL iBT Listening section to measure test takers' communicative listening ability.

It could also be argued that the TLU domains for academic speaking as a student and as an instructor might be less dissimilar than the TLU domains of academic listening as a student and as an instructor. In other words, the types of speaking required by a university instructor might not be all that different from the types of speaking required by a university student, and thus the validity argument made for the TOEFL iBT Speaking component can be more easily applied to ITA screening. However, the types of listening required by a university instructor might be more different from the types of listening required by a university student. Numerous researchers (e.g., J. D. Brown & Kondo-Brown, 2006; Gilmore, 2007; Wagner, 2014a) have described how the level of formality of a speaking context will affect the characteristics of the speaker's utterances (a more formal text will tend to have less connected speech, less slang and colloquial language, more formal organization, etc.). From this perspective, being able to understand instructors' speech (which tends to be more formal and regular) is very different from being able to understand students' speech (with much more informal speech, connected speech, and colloquial language). A more thorough analysis of the comparisons of TLU domains for speaking and listening, from the perspectives of both a student and an instructor, is clearly needed when building a validity argument supporting the use of TOEFL iBT Listening scores for ITA screening purposes.

In addition, the curriculum of the ITA 5501 class focused not only on oral language ability but also on teaching skills. Obviously, teaching skills are not measured by the TOEFL iBT. The curriculum for the teaching component of the ITA 5501 course focused on the types of language forms that instructors need to lead a class. Elder and Kim (2014), in defining the construct of teacher language proficiency, noted that teachers need to be able to do things involving classroom management and giving instructions, that involve "language forms and discourse strategies that may not be routinely used in everyday communication" (p. 2). These types of "teacher" language forms are a specific focus of the ITA 5501 curriculum. The ITA 5501 curriculum also focuses on good pedagogical techniques, appropriate nonverbal teaching behaviors, how to use classroom tools (e.g., the blackboard), and North American university classroom cultural norms. Assuming that the students in the ITA class learned these "teacher" language forms and instructional skills, and were later able to apply them to their teaching situations, their overall teaching competence might have improved dramatically, even if their gains in oral proficiency, as measured by TOEFL iBT Speaking and Listening scores, were limited. Schmidgall (2013) found that with a LSP test that focuses on teaching, it is difficult to separate language ability from the context in which the language is used. Xi (2008) found that when locally administered ITA screening exams attempted to measure oral proficiency only, the results of these exams correlated quite highly with TOEFL iBT Speaking scores. In contrast, when the locally administered ITA screening exams included nonlinguistic factors, such as teaching ability and knowledge of American classroom norms, the correlations with TOEFL iBT Speaking scores were much lower. For those ITAs who successfully completed ITA 5501, it seems likely that their ability to be successful university instructors increased, even if their improvement in speaking ability (as measured by TOEFL iBT Speaking) was modest. Oppenheim (1998) found a small but statistically significant improvement in ITAs' "teaching performance" after they completed a teaching preparation seminar than before they completed this teaching seminar. This seminar was only 24 hours long and was not a for-credit class, so it seems reasonable to believe that a 45-hour, 3-credit class devoted to teaching preparation and communicative competence would have even better results.

It also seems reasonable to believe that teaching competence can make up for some of the ITAs' oral language shortcomings. If this is the case, this would provide support for use of the "stronger" versions of language for specific-purposes tests for ITA screening; that is, including teaching competence in the scoring of the local simulated teaching tests should lead to better assessments of the test takers' ability to teach competently in the university classroom. Again, this is a

somewhat controversial notion, because of the belief that assessing teaching competence is unfair and discriminatory to the ITAs, because the domestic teaching assistants are not required to demonstrate their teaching competence (Bailey, 1985; Farnsworth, 2013; Hoekje & Williams, 1992; Saif, 2002; Schmidgall, 2013). However, this view overlooks the fact that including teaching competence on the assessment would actually be beneficial to those test takers who have learned (through experience, through the ITA preparation class, etc.) strong pedagogical skills. This might also account (at least in part) for Xi's (2007, 2008) findings that the TOEFL iBT Speaking scores correlated less strongly with those local ITA screening tests that included teaching competence in their scoring rubrics.

However, there also seems to be a language threshold below which teaching competence cannot compensate for an instructor's lack of oral English proficiency (Bailey, 1985; Elder, 1993; Elder & Kim, 2014; Halleck & Moder, 1995). Similar to how Chalhoub-Deville and Deville (2006) argued that there is a minimum threshold of English proficiency that students need to have to be able to perform in an English-medium academic context, there also seems to be a minimum threshold of English proficiency for instructors. ITAs who do meet that minimum language threshold are able to utilize the skills they learn in the ITA class (e.g., communication strategies, teaching skills, knowledge of American classroom culture, appropriate nonverbal teaching behaviors) to provide quality instruction for undergraduates, even if their oral proficiency, as measured by the TOEFL iBT, might seem comparatively low.

Limitations

A number of limitations in the study need to be acknowledged. Because of the difficulty in accessing data on former ITAs, and in getting ITAs to participate in the study, there were fewer participants than anticipated, and small sample sizes restricted the amount and types of statistical analyses employed as well as the generalizability of the results. Whereas the vast majority of the ITAs in the ITA 5501 class participated in the study, the participation rate for other ITAs was much lower and thus might suggest a sampling problem. In addition, it is problematic to compare the results of different analyses between different groups of ITA participants.

Another issue that must be addressed is related to the methodological design of the use of TOEFL iBT Speaking and Listening scores as predictors of instructional competence. Using these two variables as the only predictors runs the risk of identifying as significant the strength of these predictors, when they might actually be marginal when considered among other predictors. That is, although it seems likely that a certain level of English oral proficiency is necessary to be a successful teacher in an American college classroom, other abilities (i.e., content knowledge, pedagogical knowledge, pedagogical content knowledge) might be much better predictors of teaching competence. Unfortunately, it was impossible to include these abilities as predictor variables in the current study, and this must thus be acknowledged as a limitation of the study. In addition, it was decided only to focus on TOEFL iBT Speaking and Listening as predictors in this study, because teaching competence is usually conceptualized as involving predominantly oral proficiency and TOEFL iBT Speaking scores are used almost exclusively for ITA placement decisions. Nevertheless, it might have been informative to include TOEFL iBT Reading and Writing scores as predictor variables as well.

Using the official SFFs as a measure of teacher competence is problematic for many reasons. As noted in the literature, these undergraduate evaluations are often unreliable and are also affected by a number of construct-irrelevant factors. Because of the problematic nature of SFF evaluations as a measure of teaching competence, it was decided to include two other measures of teaching competence: students' evaluations of ITAs' teaching competence and observers' assessment of ITAs' teaching competence. The student evaluations of teaching competence used in this study were designed to improve on the SFFs yet still suffered from many of the same shortcomings. Assessing teaching competence through classroom observations avoids many of the inherent shortcomings of the SFFs and students' assessments but introduces problems of its own. Single observations of an ITA's classroom teaching provide a very limited view of an ITA's classroom teaching over the course of a semester. In addition, a single rating is almost always inferior to multiple ratings. Because of this, it was decided to use two different observers to observe and evaluate each ITA's teaching. Nevertheless, even two observational assessments present a limited view of overall teaching competence. In addition, although having the two different observers observe and evaluate two different classes of the ITA's teaching is beneficial in that two separate lessons should allow a better evaluation of a person's teaching competence, this also made it difficult to assess the reliability of the raters' scoring and the observation rubric. This lack of investigation of the reliability of the observers (and the rubric) is a noted limitation. Finally, the observation rubric used for this study was originally based on the ITA 5501 curriculum and TEACH test scoring rubric, which is a narrow operationalization of instructional competence, and thus these limitations

might affect the interpretation and generalizability of the results of this study; therefore, the results must be interpreted with caution.

One of the goals of this study was to measure the increased level of oral proficiency that ITAs exhibit after a semester of living, studying, and teaching in an English-language environment. But language learning is a slow and long process. Although the speaking and listening sections of the TOEFL iBT are reliable measurements of test takers' oral proficiency, measuring language gains made after only 3 months is difficult, and the inevitable measurement error involved in assessing language ability might mask any real gains the learners might achieve and must be acknowledged as a limitation of the study. Related to this is the number of opportunities the learners have to interact with proficient English speakers on campus. As a number of researchers have noted (e.g., Cheng & Fox, 2008; Myles & Cheng, 2003; Ranta & Meckelborg, 2013; Trice, 2004), the amount of interaction is highly variable among learners, and this study did not attempt to measure the amount of interaction by individual learners. This would be a valuable measure to investigate to determine if there is a correlation between the amount of interaction and gains in oral proficiency or even in teaching competence.

Finally, to investigate the development of English proficiency, the ITA participants took research versions of the TOEFL iBT. While these versions were equivalent to the operational TOEFL iBT, were taken under operational conditions, and were scored by trained and certified ETS raters, there is always error involved when measuring performance, and thus using difference scores as the measure of oral proficiency development is problematic (Zumbo, 1999) and is a limitation of the study. In addition, the motivational levels of the test takers might have differed from those of actual TOEFL iBT test takers. The results of these tests were not high stakes for the test takers, and thus the participants might not have been as motivated to do as well as they would have been had the results been meaningful for them (i.e., for university admission or ITA placement). In observing the test takers while they were taking the test, there was no indication that test takers were not trying—and the fact that the tests were low stakes probably served to reduce the test takers' anxiety levels. Nevertheless, this possible lack of motivation to do as well as possible on the tests needs to be acknowledged as a shortcoming.

Conclusion

Even with its many limitations, the current study yielded valuable information about the use of TOEFL iBT Speaking and Listening scores for ITA screening purposes. First, the TOEFL iBT Speaking and Listening scores correlate only weakly with the TEACH test, the teaching simulation culminating assessment for the ITA preparation class at Temple University. Although TOEFL iBT Speaking correlated weakly with the language proficiency component of the TEACH test, there was no correlation between TOEFL iBT Speaking and the teaching competence component of the TEACH test. In contrast, TOEFL iBT Listening correlated weakly, but at a statistically significant level, for both the language and teaching components of the TEACH test.

This finding was further supported by the correlation and regression analyses that investigated the relationship between the TOEFL iBT Listening scores and two of the measures of actual teaching competence. Although there was no correlation between SFF scores and the TOEFL iBT Speaking or TOEFL iBT Listening scores, this result was not especially surprising because of the numerous shortcomings of institutional end-of-semester measures of teaching competence, such as the SFFs. In contrast, the alternative (and more objective) measures of teaching competence that were used in this study (students' assessment of ITAs' teaching competence and observers' assessment of ITAs' teaching competence) did correlate with TOEFL iBT Listening scores, although they did not correlate with TOEFL iBT Speaking scores. This contrast was made even more explicit in the regression analyses. Whereas TOEFL iBT Speaking scores predicted only a negligible percentage of the teaching competence scores, TOEFL iBT Listening scores accounted for more than 20% of the observers' assessment of ITAs' teaching competence and for more than 15% of the students' assessment of ITAs' teaching competence. This provides empirical evidence for the already strong theoretical rationale of utilizing TOEFL iBT Listening scores in addition to TOEFL iBT Speaking scores for ITA placement test purposes. This is important, because currently almost all ITA programs that use large-scale proficiency test results (such as the TOEFL iBT) for initial ITA screening rely on the Speaking scores exclusively. This study provides evidence that using Listening scores as well as Speaking scores can help ITA programs make better ITA screening decisions.

Finally, the study found that ITAs do demonstrate small yet measurable increases in oral proficiency after a semester of living, studying, and (sometimes) teaching in an English-medium university setting. Although the gains in oral proficiency, as measured by the TOEFL iBT Speaking and TOEFL iBT Listening tests, were relatively small, the TOEFL iBT

Speaking score gains were statistically significant. In addition, it seems likely that the gains in oral proficiency included aspects of communicative competence that are not measured by the TOEFL iBT. This is relevant for ITA programs' screening and cut score determination as well as for curriculum development for ITA preparation courses. These results also have implications for the teaching simulation performance assessments often used as the culminating assessments for ITA preparation courses, suggesting that "stronger" versions of these English for Specific Purposes exams are warranted, where teaching competence is part of the construct being assessed.

Notes

1 In 2013, the procedure for administering SFFs changed. Starting in fall 2013, all SFFs were online, and students completed and submitted the SFFs electronically, outside of class time.

2 As mentioned, the subscales in this questionnaire were designed to mirror the different components of the classroom observation rubric. However, because it did not seem feasible to create Likert-type items in which the students would assess the instructor's ability in "delivery strategies and nonverbal communication," it was decided only to include three subscales here.

References

Abraham, R., & Plakans, B. (1988). Evaluating a screening/training program for NNS teaching assistants. *TESOL Quarterly*, *22,* 505 – 508.

Bailey, K. (1983). *Teaching in a second language: The communicative competence of non-native speaking teaching assistants* (Unpublished doctoral dissertation). University of California, Los Angeles.

Bailey, K. (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. Hauptman, R. LeBlanc, & M. Wesche (Eds.), *Second language performance testing* (pp. 153 – 180). Ottawa, ON: University of Ottawa Press.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: A lexico-grammatical analysis* (TOEFL iBT Research Report No. iBT-19). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02311.x

Brecht, R., Davidson, D., & Ginsberg, R. (1995). Predicting and measuring language gains in study abroad settings. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 37 – 66). Amsterdam, The Netherlands: John Benjamins.

Brecht, R., & Robinson, J. (1995). On the value of formal instruction in study abroad: Student reactions in context. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 317 – 334). Amsterdam, The Netherlands: John Benjamins.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT Speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29,* 91 – 108.

Briggs, S. (1994). Using performance assessment methods. In C. Madden & C. Myers (Eds.), *Discourse and performance of ITAs* (pp. 63 – 80). Alexandria, VA: TESOL.

Briggs, S., & Hofer, B. (1991). Undergraduate perceptions of ITA effectiveness. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 435 – 445). Dubuque, IA: Kendall/Hunt.

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

Brown, J. D., & Kondo-Brown, K. (2006). Introducing connected speech. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 1 – 15). Manoa, HI: National Foreign Language Resource Center.

Brown, K. (1992). American college student attitudes toward non-native instructors. *Multilingua, 11,* 249 – 265.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series Report No. 20). Princeton, NJ: Educational Testing Service.

Carrier, C., Dunham, T., Hendel, D., Smith, K., Smith, J., Solberg, J., & Tzenis, C. (1990). *Evaluation of the teaching effectiveness of international teaching assistants who participated in the teaching assistant English program*. St. Paul: University of Minnesota. Retrieved from ERIC Document Reproduction Service. (ED 175824)

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 517 – 530). Westport, CT: American Council on Education/Praeger.

Chapelle, C. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319 – 352). New York, NY: Routledge.

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.

Cheng, L., & Fox, J. (2008). Towards a better understanding of academic acculturation: Second language students in Canadian universities. *Canadian Modern Language Review, 65,* 307 – 333.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 28,* 421–442.

Collentine, J., & Freed, B. (2004). Learning context and its effects on second language acquisition. *Studies in Second Language Acquisition, 26,* 153–171.

Davis, W. (1991). International teaching assistants and cultural differences: Student evaluations of rapport, approachability, enthusiasm, and fairness. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 446–451). Dubuque, IA: Kendall/Hunt.

Derwing, T., Munro, M., & Thomson, R. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29,* 359–380.

Dick, R., & Robinson, B. (1994). Oral English proficiency requirements for ITAs in U.S. colleges and universities: An issue in speech communication. *JACA, 2,* 77–86.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series No. MS-8). Princeton, NJ: Educational Testing Service.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.

Educational Testing Service. (2005). *TOEFL iBT at a glance*. Princeton, NJ: Author.

Educational Testing Service. (2014). *TOEFL iBT test independent Speaking rubrics*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Elder, C. (1993). Language proficiency as predictor of performance in teacher education. *Melbourne Papers in Language Testing, 2*(1), 1–17.

Elder, C., & Kim, S. H. O. (2014). Assessing teachers' language proficiency. In A. Kunnan (Ed.), *Companion to language assessment* (Vol. 1, pp. 454–470). Oxford, England: Wiley-Blackwell.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. In R. Tulloh (Ed.), *IELTS research reports* (Vol. *4*, pp. 207–254). Canberra: IELTS Australia.

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., & Wagner, E. (2013, March). *Using TOEFL iBT Speaking for ITA screening: Promises and perils*. Paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Dallas, TX.

Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam, The Netherlands: John Benjamins.

Freed, B. (1998). An overview of issues and research in language learning in a study abroad setting. *Frontiers, 4,* 31–60.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40,* 97–118.

Gorsuch, G. (2003). The educational cultures of international teaching assistants and U.S. universities. *TESL-EJ: Teaching English as a Second or Foreign Language*, *7*(3), 1–26.

Gorsuch, G. (2006). Discipline-specific practica for international teaching assistants. *English for Specific Purposes, 25,* 90–108.

Gorsuch, G., Myers, C., Pickering, L., & Griffee, D. (2013). *English communication for international teaching assistants* (2nd ed.). Long Grove, IL: Waveland Press.

Graham, J. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21,* 505–521.

Halleck, G., & Moder, C. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly, 29,* 733–758.

Hinofotis, F., & Bailey, K. (1981). American undergraduates' reactions to the communication skills of foreign teaching assistants. In J. Fisher, M. Clarke, & J. Schacter (Eds.), *On TESOL '80: Building bridges: Research and practice in teaching ESL* (pp. 120–133). Washington, DC: TESOL.

Hoekje, B., & Linnell, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28,* 103–126.

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26,* 243–269.

Howell, D. (2002). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Huebner, T. (1995). The effects of overseas language programs: Report on a case study of an intensive Japanese course. In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 171–193). Amsterdam, The Netherlands: John Benjamins.

Inglis, M. (1993). The communicator style measure applied to nonnative speaking teaching assistants. *International Journal of Intercultural Relations, 17,* 89–105.

Jacobs, L., & Friedman, C. (1988). Student achievement under foreign teaching associates compared with native teaching associates. *Journal of Higher Education, 59,* 551–563.

Kang, O., Rubin, D., & Lindemann, S. (2014). Mitigating U.S. undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly, 49*, 681–706.

Kokhan, K., & Lin, C.-K. (2014). Test of English as a Foreign Language (TOEFL): Interpretation of multiple score reports for ESL placement. *Papers in Language Testing and Assessment, 3,* 1 – 23.

Lapkin, S., Hart, D., & Swain, M. (1995). A Canadian interprovincial exchange: Evaluating the linguistic impact of a three-month stay in Quebec. In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 67 – 94). Amsterdam, The Netherlands: John Benjamins.

Ling, G., Powers, D., & Adler, R. (2014). *Do TOEFL iBT scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument* (TOEFL iBT Research Report No. ETS RR-14-09). Princeton, NJ: ETS.

Llanes, A., & Munoz, C. (2009). A short stay abroad: Does it make a difference? *System, 37*, 353 – 365.

Marsh, H. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707 – 754.

Monoson, P., & Thomas, C. (1993). Oral English proficiency policies for faculty in U.S. higher education. *Review of Higher Education, 16,* 127 – 140.

Myers, C. (1994). Question-based discourse in science labs: Issues for ITAs. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 83 – 102). Alexandria, VA: TESOL.

Myles, J., & Cheng, L. (2003). The social and cultural life of non-native English speaking international graduate students at a Canadian university. *Journal of English for Academic Purposes, 2,* 247 – 263.

Nelson, G. (1992). The relationship between the use of personal, cultural examples in international teaching assistants' lectures and uncertainty reduction, student attitude, student recall, and ethnocentrism. *International Journal of Intercultural Relations, 16,* 33 – 52.

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29,* 557 – 582.

O'Loughlin, K., & Arkoudis, S. (2009, March). *Investigating IELTS exit score gains in higher education*. Paper presented at the annual meeting of the American Association for Applied Linguistics, Denver, CO.

Oppenheim, N. (1998, March). *Undergraduates' assessment of international teaching assistants' communicative competence*. Paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Seattle, WA. Retrieved from ERIC Document Reproduction Service. (ED 423783)

Papajohn, D. (2006). Standard setting for next generation TOEFL academic speaking test (TAST): Reflections on the ETS panel of international teaching assistant developers. *TESL-EJ, 10*(1), 1 – 7.

Plakans, B. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly, 31,* 95 – 119.

Plakans, B., & Abraham, R. (1990). The testing and evaluation of international teaching assistants. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (pp. 68 – 81). Washington, DC: NAFSA.

Plough, I., Briggs, S., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27,* 235 – 260.

Powers, D., & 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,* 151 – 167.

Powers, D., Schedl, M., Wilson Leung, S., & Butler, F. (1999). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16,* 399 – 425.

Ranta, L., & Meckelborg, A. (2013). How much exposure to English do international graduate students really get? Measuring language use in a naturalistic setting. *Canadian Modern Language Review, 69,* 1 – 33.

Regan, V. (1995). The acquisition of sociolinguistic native speech norms: Effects of a year abroad on second language learners of French. In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 245 – 267). Amsterdam, The Netherlands: John Benjamins.

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

Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of non-native English-speaking teaching assistants. *Research in Higher Education, 33,* 511 – 531.

Rubin, D. (2002). Help! My professor (or doctor or boss) doesn't talk English! In J. Martin, T. Nakayama, & L. Flores (Eds.), *Readings in intercultural communication: Experiences and contexts* (pp. 127 – 137). Boston, MA: McGraw-Hill.

Rubin, D., & Smith, K. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations, 14,* 337 – 353.

Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics/Revue Canadienne De Linguistique Appliquee, 5*(1 – 2), 145 – 167.

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT Listening section* (TOEFL iBT Research Report No. 8). Princeton, NJ: Educational Testing Service.

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (TOEFL iBT Research Report No. 21). Princeton, NJ: Educational Testing Service.

Schmidgall, J. (2013). *Modeling speaker proficiency, comprehensibility, and perceived competence in a language use domain* (Unpublished doctoral dissertation). University of California, Los Angeles.

Segalowitz, N., & Freed, B. (2004). Context, contact and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition, 26,* 172 – 199.

Smith, J., Myers, C., & Burkhalter, A. (1992). *Communicate: Strategies for international teaching assistants*. Englewood Cliffs, NJ: Regents/Prentice Hall.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Thomas, C., & Monoson, P. (1991). Issues related to the state-mandated English language proficiency requirements. In J. Nyquist, R. Abbott, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 382 – 392). Dubuque, IA: Kendall/Hunt.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17,* 84 – 119.

Trice, A. (2004). Mixing it up: International graduate students' social interactions with American students. *Journal of College Student Development, 45,* 671 – 687.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly, 5,* 218 – 243.

Wagner, E. (2010a). How does the use of video texts affect ESL listening test-taker performance? *Language Testing, 27,* 493 – 510.

Wagner, E. (2010b). Survey research in applied linguistics. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to second language research methods* (pp. 22 – 38). London, England: Continuum.

Wagner, E. (2012, October). *Assessing the classroom pragmatic competence of ITAs*. Symposium presentation at the annual meeting of the Language Testing Research Colloquium, Princeton, NJ.

Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly, 10*, 178 – 195.

Wagner, E. (2014a). Assessing listening. In A. Kunnan (Ed), *Companion to language assessment* (Vol. 1, pp. 47 – 63). Oxford, England: Wiley-Blackwell.

Wagner, E. (2014b). Using unscripted spoken texts to prepare L2 learners for real world listening. *TESOL Journal, 5*, 288 – 311.

Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banerjee & D. Tsagari (Eds.), *Contemporary Second Language Assessment* (pp. 438 – 463). London, England: Continuum.

Wagner, E., & Toth, P. (2014). Teaching and testing L2 Spanish listening using scripted versus unscripted texts. *Foreign Language Annals, 47,* 404 – 422.

Wagner, E., & Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high stakes tests in China, Japan, and Taiwan. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 438 – 463). Newcastle upon Tyne, England: Cambridge Scholars Publishing.

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 Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly, 4,* 318 – 351.

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for screening and setting standards for ITAs* (TOEFL iBT Research Report No. 3). Princeton, NJ: Educational Testing Service.

Xi, X., Bridgeman, B., & Wendler, C. (2013). Tests of English for academic purposes in university admissions. In A. Kunnan (Ed.), *Companion to language assessment* (Vol. 1, pp. 318 – 337). Oxford, England: Wiley-Blackwell. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02088.x

Yook, E., & Albert, R. (1999). Perceptions of international teaching assistants: The interrelatedness of intercultural training, cognition, and emotion. *Communication Education, 48,* 1 – 17.

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT* (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 269 – 304). Greenwich, CT: JAI Press.

Appendix A: International Teaching Assistant TEACH Test Evaluation Form

TA Name:_____ Rater:_____

Department:_____ Date:_____

I. PRESENTATION LANGUAGE SKILLS (circle one)

	(Inadequate)	(Adequate)	(Good)	(Excellent)
1. Comprehensibility	**1**	**2**	**3**	**4**
2. Accent and pronunciation	**1**	**2**	**3**	**4**
3. Listening comprehension	**1**	**2**	**3**	**4**
3. Fluency	**1**	**2**	**3**	**4**
5. Grammar and word choice	**1**	**2**	**3**	**4**

SCORE_____ (Out of 20)

II. TEACHING SKILLS/INTERACTIONAL SKILLS (circle one)

	(Inadequate)	(Adequate)	(Good)	(Excellent)
1. Lesson organization and implementation	**1**	**2**	**3**	**4**
2. Relevance of content and development of content	**1**	**2**	**3**	**4**
3. Interaction with students and nonverbal communication	**1**	**2**	**3**	**4**

SCORE _____ (Out of 12) **TOTAL SCORE** _____ (Out of 32)

***Rater's Comments*:** Please make comments regarding the examinee's overall performance. Your comments will serve as formative feedback to the examinee.

Appendix B: Temple University Course and Teaching Evaluations, Student Feedback Forms (SFFs)

Students respond using a 5-point Likert-type scale (i.e., *strongly agree*, *agree*, *neutral*, *disagree*, and *strongly disagree*).

Items used to assess the instructor (and student):

1. I came well prepared for the class.
2. The instructor clearly explained the educational objectives of this course.
*3. The instructor was well organized and prepared for class.
4. The instructor was conscientious in meeting class and office hour responsibilities.
5. The instructor promoted a classroom atmosphere in which I felt free to ask questions.
6. The instructor provided useful feedback about exams, projects, and assignments.
7. So far, the instructor has applied grading policies fairly.
*8. The instructor taught this course well.

Items used to assess the course:

1. The course content was consistent with the educational objectives of this course.
2. The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.
*3. I learned a great deal in this course.

*SFF items that were used in this study.

Appendix C: Student Evaluation of International Teaching Assistants' Oral Proficiency and Teaching Competence

Oral Proficiency

Fluency of Instructor's Speech

1. The instructor is fluent in English and easy to understand.
2. The instructor's speaking rate was appropriate (not too fast and not too slow).

Comprehensibility and Accent

3. I am able to understand the instructor's directions and instructions.
4. The instructor speaks clearly and comprehensibly.
*5. The instructor has a **foreign accent** when speaking.

Instructor's Grammatical and Vocabulary Accuracy

6. The instructor's grammatical and vocabulary accuracy makes it easy to understand what s/he was saying.
7. The instructor used accurate English grammar and vocabulary.

Instructor's Ability to Understand Students' Speech

8. The instructor understands the students' speech.
9. The instructor understands me when I speak to him/her.

Estimation of Instructor's Oral Proficiency

10. The instructor's oral proficiency in English is sufficient to be an instructor at this university. (1 for yes, 2 for no)
*11. The instructor should **not** be teaching this class because his/her English proficiency is too low. (1 for yes, 2 for no)

Teaching Competence

Interaction With Students

12. The instructor replies appropriately to students' questions and comments.
13. The instructor's interaction with the students is appropriate.
14. The instructor encourages and acknowledges student participation.

Instructor's Knowledge of American Classroom Culture

15. The instructor is aware of American classroom culture and norms.
16. The instructor is knowledgeable of the role of the teacher and/or the student in American university classrooms.
17. The instructor understands what is expected of a teacher in an American university classroom.

Ability to Communicate Content Information

18. The teacher communicated the content of the course clearly and coherently.
19. I was able to understand the content of the instruction.

Students' Experience With Nonnative Speakers

20. In the past, I have had an extensive amount of interaction with nonnative speakers of English.

21. I am good at understanding nonnative speakers of English.

22. Prior to this class, how many instructors have you had at Temple who were nonnative speakers of English? (1 for zero or one nonnative instructors, 2 for two nonnative instructors, 3 for three nonnative instructors, 4 for four nonnative instructors, and 5 for more than four nonnative instructors)

*Items 5 and 11 were reverse-coded in the analyses.

Appendix D: Classroom Observation Rubric

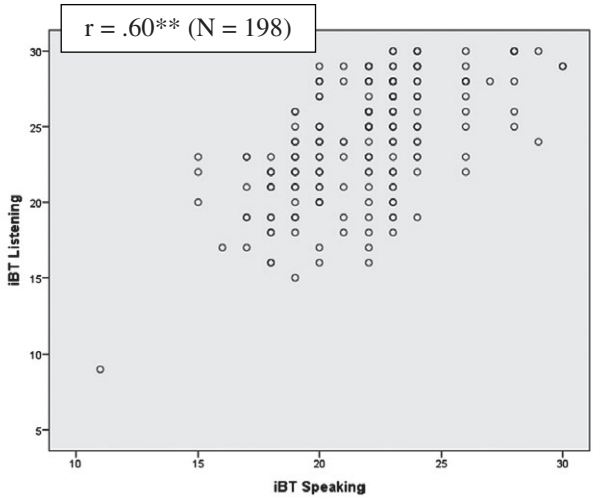
Instructor's Oral Language Proficiency

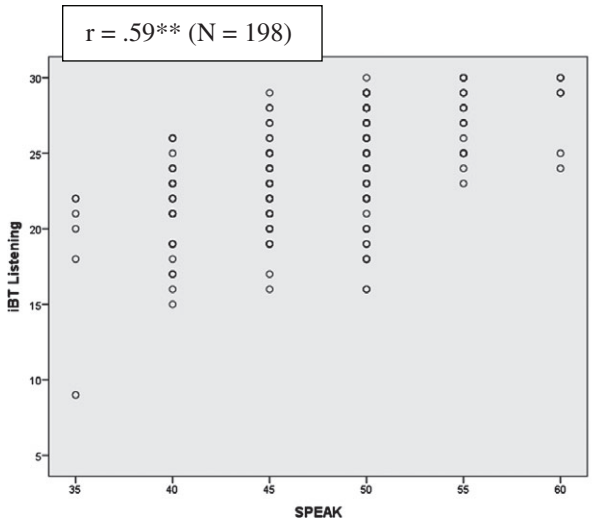
Score	Fluency	Comprehensibility and accent	Grammatical and vocabulary accuracy	Listening comprehension
4	Appropriate speech rate — not too fast and not too slow. Appropriate use and placement of pauses (mostly filled). Very fluent speech.	Speech is easily understood, and only minor instances of nontarget accent. Appropriate volume.	Very few grammatical or vocabulary errors, and these do not cause communication failures.	Instructor understands students' questions and comments at a natural speech rate. Instructor provides appropriate responses to the questions and comments.
3	Minor problems with speech rate, excessive pauses, including inappropriate placement.	Some minor nontarget aspects like stress, intonation, and pronunciation issues, although these do not make it difficult for the listener to understand the speech.	Some grammatical errors and inappropriate vocabulary use, causing only minor comprehension breakdowns.	Minor instances of the instructor not understanding students' speech. Sometimes has to ask students to repeat, or sometimes responds inappropriately, suggesting that s/he did not understand.
2	Speech rate is often inappropriate (usually too slow). Excessive pauses (usually unfilled) negatively affect listener's perception of fluency.	Many non-target-like instances of stress, intonation, and pronunciation, and these sometimes cause strain on the listener to understand.	Many grammatical errors and inappropriate vocabulary use, sometimes causing difficulties for the listener to understand the speech.	Numerous instances where the instructor did not seem to understand students' speech. Instructor often has to ask student to repeat, and occasional instances of communication breakdown.
1	Speech rate is inappropriate (usually too slow). Repeated hesitations. Unfilled pauses within clauses dominate.	Speaker primarily uses non-target-like stress, intonation, and pronunciation. Listener has to strain to understand, and patches of speech are incomprehensible.	Numerous grammatical errors and inappropriate vocabulary use. These are so numerous that often the listener cannot understand the speech.	Instructor rarely understands students' questions or comments unless repeated. Even with repetition, the instructor often does not understand students' speech.
0	Lack of fluency is dominant — making it difficult for listener to follow along.	Accent (both segmental and nonsegmental components) is so strong that speech is usually incomprehensible.	Grammatical and vocabulary errors so frequent that speech is incomprehensible.	Instructor seems to understand almost none of the students' speech.

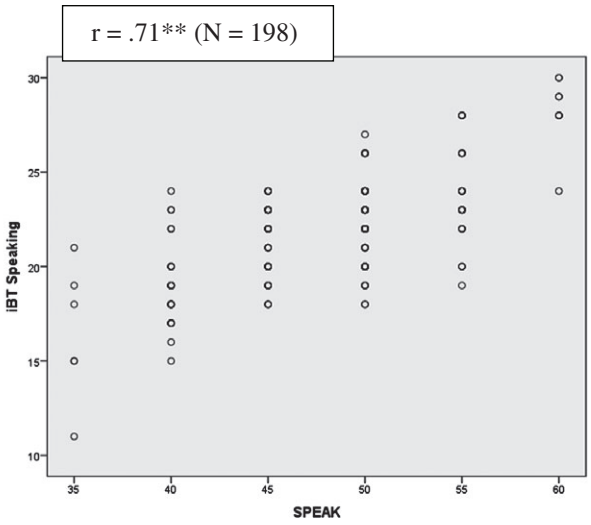
Instructor's Teaching/Interactional Skills

Score	Lesson organization and preparedness	Teacher – student interaction/ classroom culture	Delivery strategies and nonverbal communication	Communication of content information
4	Lesson shows logical development of ideas. Appropriate transitions between activities. Instructor seems prepared for all contingencies.	Instructor seems at ease in the classroom. Always interacts with students in an appropriate and expected manner.	Instructor uses clarification checks and sufficient wait time. Maintains eye contact with students, and use of body language facilitates comprehension of message.	Instructor communicates content of instruction clearly, coherently, and logically. Instructor provides examples or clarifications. Students grasp the focus of instruction.
3	Mostly logical lesson organization and development, although sometimes lacking coherence. Transitions usually appropriate. Instructor seems to have prepared adequately.	Instructor usually interacts appropriately with students, teaching style is appropriate, although some cases of not replying appropriately to student questions or concerns.	Instructor usually uses clarification checks and mostly provides sufficient wait time. Nonverbal communication somewhat lacking, or inappropriate at times.	Instructor's communication of content is mostly coherent, but sometimes lacks use of supporting evidence or clarity of expression, causing only minor difficulties for students to understand content.
2	Lesson organization and development often lacks coherence, with few logical transitions between activities. Evidence of a lack of preparedness.	Some inappropriate instructor behavior/interactions/teaching style. Instructor might ignore some student questions or comments, or respond inappropriately.	Instructor only occasionally uses clarification checks and often does not provide sufficient wait time. Nonverbal communication often inappropriate or lacking, but overall facilitates comprehension.	Instructor's communication of content is somewhat coherent and nnnnnn expressions. Students sometimes have difficulty understanding content.
1	Although some attempts at lesson organization, the instructor seems unprepared, and the lesson lacks coherence.	Instructor's teaching style and overall interactions with students is inappropriate, causing noticeable student unease.	Instructor almost never uses clarification checks, does not provide sufficient wait time, and nonverbal communication does little to facilitate comprehension.	Instructor's communication lacks coherence and is often illogical. Students have repeated difficulties understanding content.
0	No lesson organization, and the instructor seems to have done no preparation whatsoever for the lesson.	Instructor's teaching style and interactions with students is totally inappropriate in an American college classroom.	Instructor's delivery strategies are inappropriate, and nonverbal communication distracts students and hinders comprehension.	Incoherent communication. Students understand virtually none of the content information provided by the instructor.

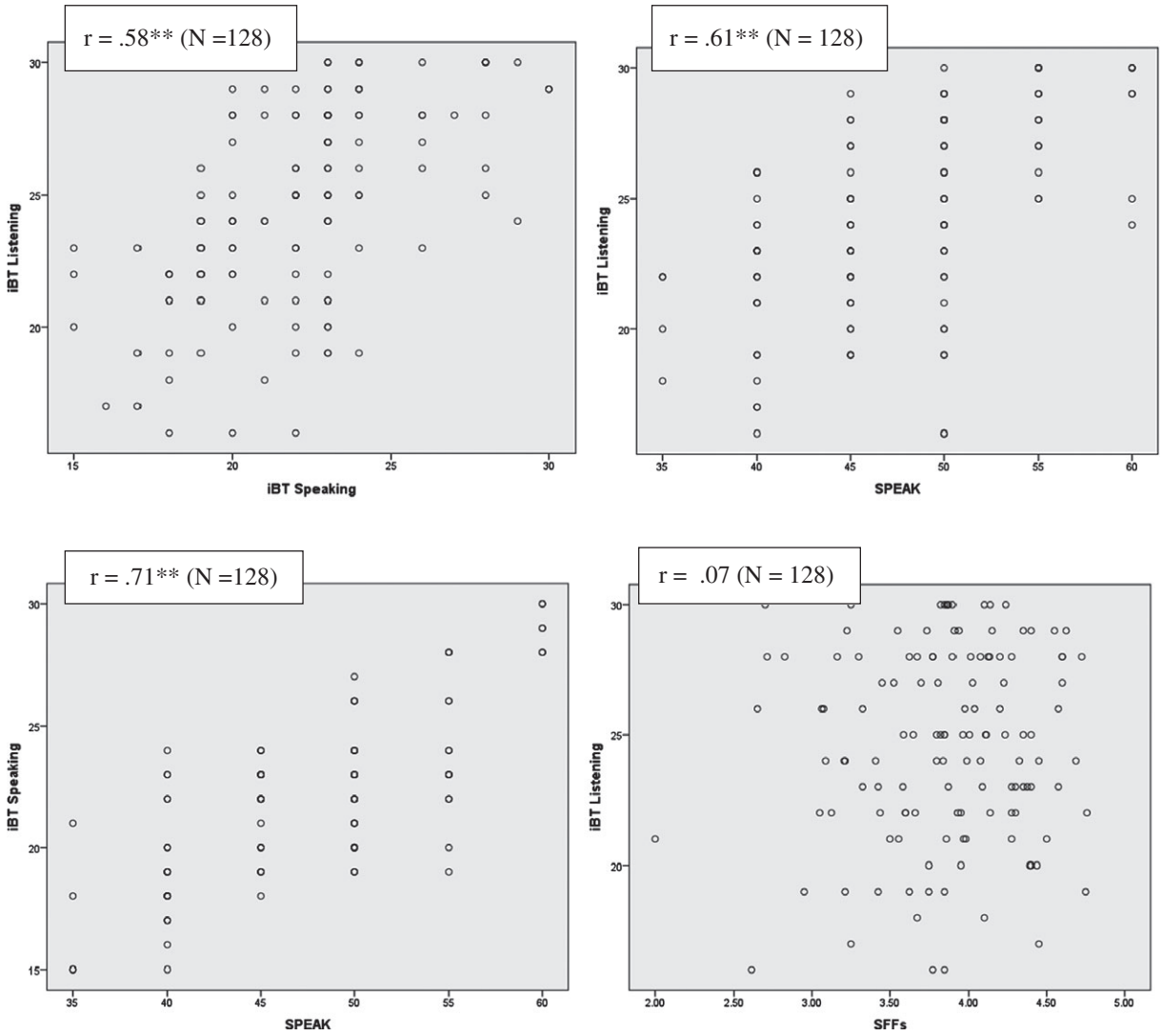
Appendix E: Scatterplots of Correlations for Research Question 1, TOEFL iBT Listening, TOEFL iBT Speaking, the SPEAK Test



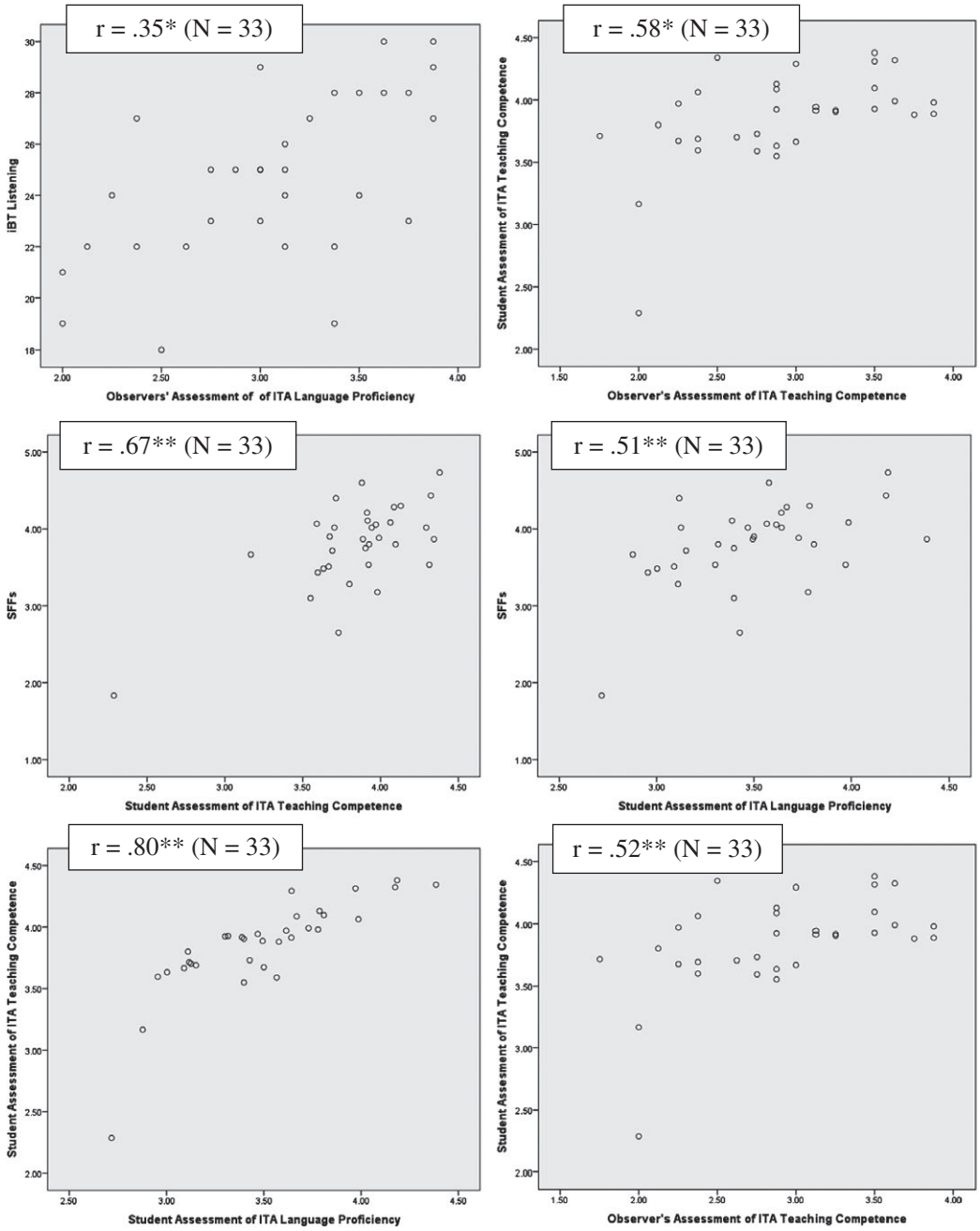




Appendix F: Scatterplots of Correlations for Research Question 2, TOEFL iBT Listening, TOEFL iBT Speaking, the SPEAK Test, and SFF Scores



Appendix G: Scatterplots of Correlations for Research Question 2, TOEFL iBT Listening, TOEFL iBT Speaking, Student Feedback Forms, Student Assessment of International Teaching Assistant Teaching Competence, Student Assessment of International Teaching Assistant Language Proficiency, Observer's Assessment of International Teaching Assistant Teaching Competence, Observer's Assessment of International Teaching Assistant Language Proficiency



Suggested citation:

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. http://dx.doi.org/10.1002/ets2.12104

Action Editor: Gary Ockey

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