ESTABLISHING THE VALIDITY OF THE TASK-BASED ENGLISH SPEAKING TEST (TBEST) FOR INTERNATIONAL TEACHING ASSISTANTS

by

Autumn Song Witt

A Dissertation Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY DOCTORAL PROGRAM IN
SECOND LANGUAGE ACQUISITION AND TEACHING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2010

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by Autumn Witt

entitled Establishing the Validity of the Task-Based English Speaking Test (TBEST) for International Teaching Assistants

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy

____ Date: 03/30/2010
Jun Liu

____ Date: 03/30/2010
Jonathan Reinhardt

____ Date: 03/30/2010
Darrell Sabers

____ Date:

____ Date:

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____ Date: 03/30/2010
Dissertation Director: Jun Liu

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Autumn Witt

ACKNOWLEDGEMENTS

Thank you…

Jun Liu, for your support and guidance, you are the quintessential teacher-researcher. I learned so much as your research assistant, but mostly that there is always a lesson to be learned and then shared through publication. Your passion for improving the field of TESOL is infectious!

Jon Reinhardt, for sharing your own dissertation experience with me and discussing my research every step of the way. Thank you for always looking for the counter-argument, and hopefully keeping my research balanced.

Darrell Sabers, for your keen eye for detail. This dissertation began as an assignment in your class, and I thoroughly enjoyed following this research path with you.

To my TBEST co-administrator, Karen Barto-Sisamout, for making research fun! The past two years would not have been the same without you and I am grateful for all those testing Saturdays that we were able to spend together.

Jacob, for being such a supportive husband and editing partner, and taking great care of Elliot, so I could write. I truly could not have done this without you. Thank you Elliot for giving me plenty of distractions from writing and helping me put everything into perspective. Thank you little Witt-on-the-way for making your presence known when I tried to sleep at night, motivating me to be productive in those early, quiet hours.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation follows an assessment tool from initial design and implementation to validity analysis. The specialized variables of this study are the population: international teaching assistants and the purpose: spoken assessment as a hiring prerequisite. However, the process can easily be applied to other populations and assessment goals.

While evaluating the TBEST and TAST (TOEFL Academic Speaking Test), I search for a preponderance of evidence for assessment validity that indicate the most appropriate tool for evaluating potential ITAs. The specific evidences of assessment validity that are examined are:

1. Evidence of Domain (Content) Validity: Which test, the TBEST or the TAST most closely measures the actual skills needed to be an ITA?
2. Evidence of Predictive Criterion Validity: Which test, the TBEST or the TAST, is more valid in predicting ITA teaching success based on end of semester student evaluation (TCEs)?

Following the analyses of these points of evidence, the results of a follow-up survey of ITA impressions about the ITA training and evaluating process are reviewed. Reviewing the results of this survey places the language assessment and hiring process recommendations within its larger context, directing attention toward suggestions for improvement of the ITA training and evaluating procedures.

Over the course of 18 months, 335 ITAs were assessed using the TBEST. 193 ITAs took the TAST prior to taking the TBEST, and those scores are used for correlation analysis. 119 ITAs participated in a follow-up survey about their ITA experience.

Analysis of domain validity shows that the TBEST is better suited for assessing ITAs than the TAST due to specialized assessment content not present on the more generic TAST. The TBEST is marginally better at predicting teaching success, though the results were statistically insignificant and recommendations are made for a follow-up study. Post-hoc analysis of the discriminative utility of both tests show that the TBEST results show more useful shades of distinction between candidates while the TAST results place the majority of students in a ‘fair’ category which requires secondary interviews to assess teaching ability.

CHAPTER 1

INTRODUCTION

1.1 International Teaching Assistants

International students are a growing student population of graduate and professional programs in the United States. Growing by 8% to 671,616, the all time peak, in 2008-2009 ("Open Doors," 2009). With that increase, the number of international teaching assistants (ITAs) providing undergraduate instruction at American universities has also grown. Hiring ITAs provides universities with the benefit of affordable undergraduate instruction, and also attracts top international scholars to their programs through ITA funding.

As state and federal funding of education has fallen, universities have come to depend on teaching assistants. At the University of Arizona (UA), of the total employees there are equal percentages of faculty and graduate assistants; there are 19% faculty and 19% graduate assistants ("UA Fact Book," 2009, p. 40). In 2008-09 there were 2,856 graduate assistants and 869 were international graduate assistants (p. 42). At the UA, 60% of ITAs are from India and China, and the remaining 40% come from all other regions of the world. After growing steadily, the most popular programs are Business, Engineering, Math, and Computer Science ("Open Doors," 2009) and these are also the top departments, which hire ITAs.

ITAs have been the focus of TESOL and applied linguistics researchers since the late 1970s. Common themes of ITA research are student attitudes toward learning from foreign teachers (Bailey, 1983; Mestenhauser, 1981; Numrich, 1993; Rubin, 1992; Tyler, 1995; Liu, 1999; Fitch & Morgan, 2003; Finder, 2005), differences from American TAs (Ronkowski, 1987; Tyler, 1992; Williams, 1992; Douglas & Selinker, 1994), and the effectiveness of ITAs (Norris, 1991; Jacobs & Friedman, 1988; Kulik et al. 1985). Additionally, some researchers have focused on training recommendations (Davis, 1987; Myers, 1994; Myers & Plakans, 1991; Rounds, 1987; Smith 1993; Reinhardt, 2007) and methods of assessing ITA fluency (Yule & Hoffman, 1990; Dunn & Constantinides, 1991, Gallego, 1990).

In this dissertation, I enter the ITA discussion with those paths already well established and some debates continuing on. I do not focus on most of these issues, but rather focus on the language assessment tools that universities use to verify the eligibility of ITAs to teach undergraduates. I specifically evaluate the validity of a newly adapted test of spoken English ability, the Task-Based English Speaking Test (TBEST), which is used at the University of Arizona. I also survey ITAs at the University of Arizona (UA) about their overall feelings about hiring, training, and evaluating procedures in order to make formative recommendations for improving this process and undergraduate instruction as well.

Most importantly, I document the decision-making process of creating a standardized assessment tool from design and implementation to validity analysis. By detailing this process, this dissertation will be useful to other teachers and administrators even if the assessment goals and assessment populations differ from this context. This process takes into account what Falsgraf (2009) calls the ecology of assessment, an orientation toward the “mutually dependent relationships . . . of the needs of students, curricular changes, and community demands for accountability" (p. 495). John Norris (2000) also advocates a holistic view of assessment that recommends each instrument be "accompanied by a statement of intended use that takes into account the 'who', 'what', 'why', and 'so what/impact' of assessment."

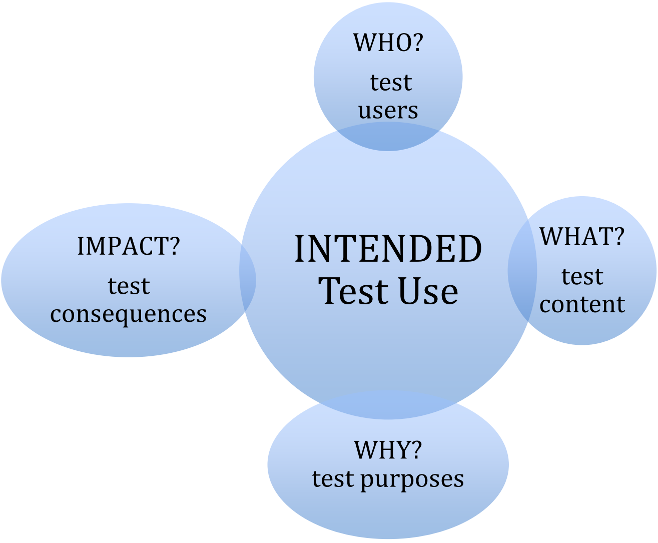


Figure 1.1. *Components of Intended Test Use (adapted from Norris, 2000).*

1.2 Importance of Language Testing

While ITAs are a consistent and reliable pool of teaching applicants, language testing is important for many reasons. First, undergraduates are right to expect competent instruction. University tuition, in both public and private institutions, increases each year and the university has an obligation to continue providing the educational services that they promise. The public still values higher education, and so they will continue paying for that advanced education (and hiring those who have it) as long as rigorous standards are maintained. Even with the rise in for-profit higher education, most students will not likely be lured away to the for-profit, online universities. Such proprietary universities' students have the highest student-loan default rates in the country (Wilson, 2010, p. 8), perhaps due to public perception about the academic quality of for-profit degree programs. For now at least, there is still value in pursuing a traditional, though perhaps not face-to-face, education, but that value and quality needs to be protected with rigorous hiring standards.

Furthermore, if ITAs are the primarily instructors in their courses, they likely teach introductory classes, so miscommunications and misunderstandings in those foundational classes could effect their students' success in the program as a whole, or even dissuade students from pursuing a degree in that field. In other classes, ITAs may also be the primary point of contact for new students, even if they are not the primary instructor. In some classes at the UA, an expert professor leads a lecture-based class of 100+ students, and then students interact with the ITA that is assigned to their section for all of their individual questions, to go over homework, and to prepare for exams. Departments can't afford to have unqualified instructors having that much contact with students.

Secondly, even though English is widely expanding as a lingua franca in many countries in the world, there is no standardized English instruction. An English ability exam is necessary because countries vary in the amount of English that is taught to students, from being the language of instruction in Hong Kong, to essentially being an elective in Japan. Since there are differing English language education requirements in different countries, American universities cannot rely on country specific English-proficiency exams or secondary language courses. There must be a university standard that is applied to all ITAs to guarantee an English ability baseline regardless of country of origin.

Specifically, spoken English competency is especially important to test. English language learners have varying opportunities to use spoken English in their home countries, so their language study may have been primarily of written English. Teaching a class requires spoken, spontaneous language ability, which is more complicated than written, asynchronous communication. It is equally important that spoken English assessment for ITAs focuses on the type of language that will be necessary for communication as an instructor. Simply testing ITAs on self-introductions and brief prepared speeches cannot predict how well they will be able to present an extended lesson, answer students' spontaneous questions, and conduct classroom management tasks.

Finally, assigning ITAs to teach who do not have the necessary language ability not only affects undergraduates, but will affect the studies and time to degree of those ITAs, as they will have to spend enormous amounts of time preparing for class and will likely meet greater resistance from their students. The English language abilities of international graduate students needed to study at the graduate level and to teach undergraduates are different. For some ITAs, it may be better for them to focus on their own studies rather than to struggle through a teaching position that compromises their own research.

1.3 Dangers of Language Assessment

Even though it is essential to test the language ability of ITAs, there are some dangers of testing that need to be acknowledged and confronted. First, there are errors of the first and second type that can emerge in any decision process. Type I errors, also called false positives, could lead to passing unqualified candidates. Type II errors, also called false negatives, could result in rejecting qualified candidates.

Table 1.1

Type I and Type II Errors in ITA Language Assessment

		Actual Candidate Condition	
		Qualified	**Not Qualified**
Test Result	**Judged “Qualified to Teach”**	True Positive	False Positive (Type I Error)
	Judged “Not Qualified to Teach”	False Negative (Type II Error)	True Negative

The consequences to Type I and Type II errors are not equal, so an institution should weigh which is most egregious to their learning community when setting cut scores or hiring requirements. What ever cut score or qualification baseline is set, there will always be a cluster of candidates that fall right on the border, and in some cases it may better to err on the side of disqualifying a few eligible candidates, at least temporarily, in order to protect the overall educational integrity of the department or school. In other cases, there may be more of an acceptance of the necessary learning curve in being a TA, and so a slightly lower cut score will be acceptable to make sure that all qualified candidates make it through screening.

In the case of errors of the first type, passing unqualified ITAs propagates negative stereotypes against the majority of ITAs, who are probably very well qualified, and more than competent to teach undergraduates. In Fitch and Morgan's (2003) narrative study, over 900 students were interviewed and approximately 70% of the stories they shared were of negative ITA experiences, but there is no mention of how many ITAs, or what percentage of the ITAs this is thought to represent (p. 303). At this particular university, the atmosphere was so hostile toward ITAs that a candidate for student body representative ran on the platform: Candidate Sloan will solve 'The ITA Problem' (p. 297). This would have been a toxic environment in which to serve as an ITA, and while students should take responsibility for their own learning (and prejudices), some of the blame may fall on an ineffective screening process for potential ITAs.

When under-qualified ITAs make it to the classroom, the experience resonates with students and prejudices all their future ITA interactions. Even though most ITAs are very capable, when a minority has significant communication difficulties, the whole ITA population becomes suspect, and competent ITAs get painted as the exception to the rule (Fitch & Morgan, 2003, p. 305). As I discuss later, no level of ITA screening can overcome some students' preconceived prejudices, but departments need to take the appropriate steps to establish an appropriate baseline of English language ability. Furthermore, language ability alone does not predict effective teaching ability, so departments should provide comprehensive training to equip ITAs to be effective instructors.

However, for some institutions, it may be more detrimental to not pass qualified ITAs, so false negatives would be a more problematic error and the cut scores or requirements should be adjusted to limit these occurrences. When a strict ITA policy turns away desirable international graduate students, programs may be more interested in making sure their ITAs pass the screening. If the departments' emphasis is on recruiting top graduate students (for both international and American students) TA positions may be merely a means to fund graduate study, with a tolerance for potentially lower satisfaction among undergraduate students.

Furthermore, some institutions prioritize the teacher training process, they understand that most graduate students will come in to the TA position with limited experience, and so departments plan extensive training. With a well-developed ITA training program and mentor support system, borderline ITAs may be able to make great improvements over the course of the semester and overcome initial language challenges. Neither of these conditions assume that no screening takes place, just that the cut scores are set with more intention to pass all qualified candidates, even if a few true negatives pass through with the population of false negatives.

In the case of errors of the second type, sometimes qualified candidates have a bad testing day, and therefore perform below their ability. These candidates should have no problem re-testing and passing on a second evaluation. However, the answer to borderline cases is not the implementation of a different assessment tool (such as an interview), because the institution needs to have reliable standards. Interviews, for all their other strengths, have too many complicating variables that reduce the overall reliability of the whole[1] process.

In either case, a lot of variables contribute to the effectiveness of a teacher. Language ability is just one variable, but personality, personal investment and interest in the topic, instructors' conflicting time commitments, rapport with students, topic knowledge, and many others also contribute. No screening test could possibility account for all these conditions and predict which candidates will succeed or fail.

In addition to Type I and Type II errors, the second type of danger is the misapplication of the assessment tool. Every test is designed with an intended purpose. Misapplying an exam gives the impression of valid results that may not actually measure the necessary or intended skills or abilities. Once a test is developed, a time- and cost-intensive process, it is tempting to use that tool for as many applications as possible. In the case of ITAs, using the same test for admittance and employment screening. However, the baseline language requirements to study in a class and to lead a class are very different, and unless the assessment tool has content that matches both of those domains, then one of the score results will be invalid for its particular application.

Misapplication of an assessment is similar to a Type III error; the hypothesis may be proved correct, but the wrong question was asked in the first place. In the case of ITA assessment, this would be like searching for predictive criterion correlations, but ignoring domain validity (Xi, 2008). While there may be correlations that emerge, if the

[1] Oral proficiency interviews (OPIs) are discussed at length in section 2.3.3.

test is not actually testing skills in the target domain, the test results may not be relevant.

1.4 Purpose of this Dissertation

In view of the above, in this dissertation I try to find a balance between the major stakeholders in the ITA hiring process at the UA to find the most effective and valid means of making hiring decisions. The graduate college needs to screen a high volume of ITAs in a short period of time and verify to the Board of Regents the English ability of all ITAs that are hired. Over 50 departments across campus depend on the graduate college to quickly evaluate potential ITAs so that they know how they can fund their international graduate students and also to know how they will staff their undergraduate general education courses. International graduate students wait anxiously to know how they will fund their graduate studies. And finally, undergraduate students expect to be able to understand their instructors. With these varied priorities, it is important that enough ITAs are eligible to teach, but they have to be truly qualified.

The evaluation population in this situation is specialized, students studying at the graduate level in a second language, so their overall language ability is much higher than among the general English language learner population. However, their language ability needs to be evaluated specifically for potential teaching ability, not just receptive ability necessary to participate as a student in English medium courses.

1.4.1 Research Questions

After a preliminary discussion in Chapter 2 of the possible ITA assessment tools, I conclude that the TBEST (Task-Based English Speaking Test) and TAST (the TOEFL

Academic Speaking Test) are the best options for screening potential ITAs. As I evaluate the TBEST and TAST, I search for a preponderance of evidence for assessment validity that indicate the most appropriate tool for evaluating potential ITAs. In this dissertation, the specific evidences of assessment validity that I examine are:

1. Evidence of Domain (Content) Validity: Which test, the TBEST or the TAST most closely measures the actual skills needed to be an ITA?
2. Evidence of Predictive Criterion Validity: Which test, the TBEST or the TAST, is more valid in predicting ITA teaching success based on end of semester student evaluation (TCEs)?

Following the analyses of these points of evidence I review the results of a follow-up survey of ITA impressions about the ITA training and evaluating process. Reviewing the results of this survey allow me to place the language assessment and hiring process within its larger context and direct attention toward suggestions for improvement of the overall ITA system.

1.5 Chapter Overview

1.5.1 Chapter 2: Literature review. Every assessment population is unique, and so I begin Chapter 2 with a survey of ITA stereotypes among undergraduates to help set the employment context. I then present an overview of the hiring, training, and evaluating procedures at universities across the country and also at the University of Arizona, the site of the present study.

Next, since the overarching goal of this dissertation is to document the assessment development process, I outline possible directions that can be pursued in developing a test. I begin by describing potential test functions: norm-referenced and domain-referenced assessment. Then I summarize different implementations of tests, summative, formative, or interim assessment. Finally, I outline potential test formats that are used to test spoken English ability: single skill assessment, oral proficiency interviews, integrated-skills assessment, and task-based assessment.

Since I am reviewing a new assessment tool (TBEST) that is based on task-based assessment, I provide additional background information on this topic. I describe the theoretical argument for task-based assessment, and also review potential definitions of a task before providing the definition of a task that is implemented on the TBEST.

1.5.2 Chapter 3: Methods and materials. In Chapter 3, I establish my research paradigm of mixed-methods research. I define my own role as the researcher and describe the participants in this study. I also outline my analyses procedures, which include parallel test form and domain analysis of the TBEST and TAST, predictive criterion analysis of TBEST and TAST scores with TCE (student evaluations) results as the criterion data, and qualitative contextualization of ITA survey results. Finally, I thoroughly describe the four data collection instruments utilized in this study: the TBEST and TAST language tests, the TCE end-of-semester student evaluations, and the follow-up ITA survey.

The TBEST was initially created in China to assess the English ability of primary and secondary school students. The testing concept was proposed for screening ITAs at the UA, and a committee lead by Jun Liu redesigned the test content for the ITA context. The TAST is the speaking portion of the TOEFL iBT, which was designed by ETS. The TCE is the standard method of evaluating all instructors at the UA and is supported by the Office of Institutional Research and Planning Support. I created the follow-up survey. It is a short (13 questions), anonymous, online survey, which was administered through surveymonkey.com.

1.5.3 Chapter 4: Analysis of results. Chapter 4 contains the analysis of all three areas of this study. The first research question concerns a domain analysis of the TBEST and the TAST. I compare both tests to a list of tasks that were initially defined by the UA TBEST development committee and then confirmed by the ITA follow-up survey. The tasks that I discuss are: description, explanation, paraphrase, justification, instruction, and classroom management.

Regarding the second research question, I review the evidence of predictive criterion correlation between the TBEST and TAST and the TCEs. These results are tempered by the small population available (n=11), so only emerging trends are discussed with a recommendation for a larger follow-up study.

Thirdly, I discuss the qualitative contextualization provided by the results of the ITA follow-up survey. Specifically, I report the comments about ITA training & evaluating, procedural issues of the hiring process, and World Englishes. The survey results equip me to make recommendations for improving the ITA hiring process, as well as ITA training, and course evaluation.

1.5.4 Chapter 5: Discussion. Chapter 5 concludes this dissertation. I begin with a brief summary of the results to my two research questions and the ITA survey. Next, I discuss the implications of this research for ITA training and evaluation and also for future research on this topic. I conclude with my personal reflections of this research process.

CHAPTER 2

LITERATURE REVIEW

Through this dissertation process, the core activity is following an assessment tool from design to implementation to validity analysis. The specialized variables of this study are the population: international teaching assistants and the purpose: spoken assessment as a hiring prerequisite. However, the process that I describe can easily be applied to other populations and assessment goals.

In this literature review, I take the reader though research available at the various decision points along the assessment development path. I begin by describing the test sample population, including ITA hiring, training, and evaluating processes at various universities across the United States. My review of ITA issues is not comprehensive as ITAs are a widely studied population[2]; I focus just on what is relevant in this study. Following the ITA review, I focus more specifically on assessment[3], reviewing the issues of test function and test format.

2.1 Describing the Research Population: International Teaching Assistants

International student enrollment in the United States increased 8% in 2008-2009 to the highest levels ever, to 671,616 students ("Open Doors," 2009). This is a 14%

[2] For an introduction to many ITA issues, see Briggs, S, et al. (1990). *The international teaching assistant: An annotated critical bibliography.*

[3] In the UK, there is terminology distinction with *assessment* referring to student work and *evaluation* referring to overall course or course delivery judgments (Taras, 2005, p. 467). However, in American academia these terms are more generic with merely connotative differences that may vary from person to person. Reynolds et al. (2006) give preliminary definitions of those connotations, that *testing* has the most negative connotations, *measurement* is also rigid and sterile, leaving *assessment* as the most neutral term (p. 3), therefore, I will primarily use the most neutral term, *assessment* and *test* will refer to the actual documents or tools.

increase over the previous peak in 2003-2003. Of those international enrollees, 46.8%, or 283,329, are graduate students. Graduate students are typically funded through TA positions, and though this is not specified in the Open Doors report, the report finds that 22.7% of the total international student population (152,457) are primarily funded through U.S. universities ("Open Doors," 2009). At the University of Arizona (UA), international undergraduates are not eligible for any university scholarships, and the primary funding for graduate students is through teaching assistantships or research assistantships. It is safe, therefore, to estimate that there are 75,000-100,000 international teaching assistants in the U.S. each year.

ITAs are a steady employment population. At the UA, ITAs composed 11.4% of the graduate TA population in 2001, and in 2008 that percentage grew slightly to 12.7%. Since this group is sizable and consistently used, study of ITA policies and procedures is warranted. Also, the face of the university is reflecting a more global society. ITAs have become commonplace, teaching undergraduate courses since the late 1970s (Bailey, 1983), incoming student classes have more international students ("Open Doors," 2009), and students have more international opportunities, with Study Abroad rates among undergraduates increasing 150% in the past decade ("Open Doors," 2009). With all this international and intercultural exposure, presumably there should be less prejudice against ITAs, but ITAs continue to face negative stereotypes in the classroom and challenges from parents, students and legislators (Finder, 2005).

2.1.1 ITA stereotypes among undergraduates. International teaching assistants are widely utilized for teaching undergraduate courses. The University Office of Institutional Research and Planning Support at the UA (www.oirps.arizona.edu) found that first-year students are the most critical population when giving end-of-course evaluations, especially when rating novice teachers ("TCE Guide," p. 20). For ITAs, the teaching challenge is even greater due to widespread negative stereotypes about ITA competency. In Fitch & Morgan's (2003) narrative study of ITA identity construction, nearly 900 undergraduates participated in telling stories about their interactions with ITAs. The majority of the stories (71%) were negative (p. 303). Furthermore, the researchers found consistent themes emerging in the construction of ITA identity:

1. Linguistic misunderstanding is the central complaint, even when the ITA simply did not speak 'the right kind of English,' when the ITA was a native English speaker of a different English variety.
2. The misunderstandings were also manifested in pedagogical complications in classroom procedures like calling names during roll call, and being unable to manage students' bad behavior. Students cast themselves as innocent victims of the ITAs' lack of classroom control.
3. In their disenchantment, the students directed bitterness toward the university, suggesting that by hiring ITAs who speak poor English the university stands to gain financially as students are forced to re-take classes. (Fitch & Morgan, 2003, pp. 304-305)

These themes help to define the difficulty that ITAs face in the classroom. Even native English speaking ITAs are criticized when they do not have a preferred accent, so ITAs who have learned English as a second or additional language face an uphill battle in that their attempts at academic English mastery will most likely not be recognized by their students. Secondly, students use simple mistakes like mispronouncing names as a scapegoat for their own misbehavior, rather than seeing themselves as equal contributors to the overall cohesion or dysfunction of the classroom atmosphere. Finally, university administrators should be especially aware of the cynical suggestion that the university stands to benefit financially if ITAs are incompetent communicators or instructors.

In the uncommon positive tale, ITAs are characterized with demeaning adjectives like 'cute' and 'little' that undermine the ITAs' classroom authority. Students even tell stories about so-called ITAs who were actually international full faculty members (p. 301). At Syracuse University, students were quoted as avoiding any teacher with a foreign sounding name, even if those professors are simply Japanese-American or Indian-American (Walters, 1993, p. 12). These trends hint at underlying xenophobia possibly at the root of complaints against ITAs as much as actual communication difficulties.

Rubin (1992) conducted a series of matched guise studies on the effects of non-language factors (such as ethnicity) on undergraduate perception of spoken comprehensibility. In the study, undergraduates (n=62) were shown pictures of a supposed instructor, either Chinese or Caucasian, and then were asked to perform a cloze listening test, filing in every seventh word, of a recorded lecture text. Based just on visual differences, the students identified lower comprehension and perceptions of accent (p. 519), when no difference existed in the recorded sample. This study, as well as others that Rubin conducted as follow-up, confirm that communication is a reciprocal process and if “communication outcomes are poor in classes taught by [ITAs] perhaps responsibility ought to be shared among [ITAs] and American undergraduate student body” (p. 512). In this instance, even vigorous pronunciation practice on the part of the ITAs could not have overcome the ethnic prejudices of the undergraduates.

2.1.2 Hiring procedures. Following state and university requirements for language screening of ITAs, most universities have hiring procedures that include some measure of spoken English ability to supplement general language ability that must be verified for graduate student admission. TSE, SPEAK, or OPIs are the most common methods with more recent iterations including more communicative and pedagogical concerns in the assessment content (Saif, 2002). In the following section, I summarize representative samples of current methods of ITA hiring assessment at American universities.

The hiring procedures for ITAs vary across universities in the United States. At Fitch and Morgan’s university, ITAs are assessed reading English aloud, performing a teaching demo, and during student-teacher role-play simulations. The assessment is recorded, ostensibly for the ITA to have the opportunity to review and learn from. However, as the researchers note, more often, this tape is used as evidence of ITA (in)competence after student complaints (2003, p. 300).

Another example of ITA hiring procedure at University B includes multiple options for evaluation; the evaluation committee uses standardized test scores, but also recordings of academic presentations or classroom teaching, or extended oral proficiency interviews (OPIs) ("Procedures," p. 2). This method is flexible to individual ITA's strengths and experiences, so examiners may be able to obtain authentic speaking samples on which to base their hiring decisions.

A standardized approach can be seen at University C. At this school, candidates must pass initial language requirements for admissions, the IELTS or TOEFL, and then specific speaking assessment benchmarks on those exams must be reached for the student to be considered for an ITA position. Candidates whose scores are below the benchmarks can request an appeal interview ("Conditions," p. 1). Numbers are not available of how many ITAs are certified through interviews as opposed to through the high standardized test requirements. Following the certification process, all ITAs participate in an extensive ITA training course ("Consulting FAQ," pp. 3-4).

At the UA, all potential ITAs take the TBEST, the task-based assessment tool described in this dissertation. The tasks are specifically designed with English spoken competence for teaching in mind. Students are rated on a 10-point scale, and departments have the flexibility to set their own cut scores. The majority of departments require a 7 or higher to designate full teaching duties, and a score of 6 to designate limited student contact, such as in a lab or as a grader. Additionally, ITAs are eligible to teach their native language (i.e., Chinese or Arabic language courses) with an English speaking score of 6. International students cannot be hired as ITAs with a score of 5 or lower, and may retake the TBEST after a semester during which they are encouraged to seek out language training.

In this section, I examined three representative samples of ITA hiring procedures including role-play and OPI, as well as standardized tests, like the generic TOEFL and the task-based TBEST. Additional discussion of possible problems and merits to these methods are discussed in section *2.3 Test Formats.* In the following section, I discuss the ITA training procedures at various universities.

2.1.3 Training. State legislatures, parents and students apply increasing pressure to ensure that ITAs are qualified and prepared to provide undergraduate instruction. It is important then, that once ITAs are hired, teaching assistance is provided. However, as Lewis (1997) rightly asks, how do programs motivate ITAs to teach effectively and conscientiously if they are not ultimately interested in a teaching career (p. 11)? Luckily, Smith found that many ITAs do have professional goals that are benefited by ITA experience (Smith 1993, Smith & Simpson 1993) which can be harnessed to motivate ITAs through the training process. In the following section, I describe ITA training programs at various universities, with a focus on representing diverse training programs. Following the survey of existing programs I review the training suggestions made by ITA researchers.

At University D, international graduate students who are interested in becoming TAs participate in a 2-week ITA training ("ITA Training"). The training includes daily linguistic training, addressing pronunciation and communication strategies. Secondly, the culture of the student body is discussed, including student demographics and admittance procedures as well as presentations on ITA experiences and student expectations. Finally, potential ITAs are given pedagogy suggestions for leading discussions, student-teacher conferences, and establishing positive student rapport. Classroom management, cheating, and unacceptable student behavior are also discussed. At the end of this training, students take the SPEAK test a final time to confirm their eligibility to teach.

At University E, ITAs have a brief, 6-day training session prior to language testing and hiring screening. Students who do not meet the testing benchmarks may enroll in a well-developed sequence of courses to guide their professional development prior to their re-screening and ITA appointment ("ITAP"). Potential ITAs are given the opportunity to enroll in 2-credit courses in Practicum, Advanced Practicum, Oral Communication, or Cross-Cultural Dynamics. There is extensive support through discussion, modeling, mentoring, and classroom observation throughout these courses to help equip students to remediate language difficulties and to be pedagogically and culturally acculturated to campus expectations before their appointment. Some ITAs pass the language requirement, but are still required to attend some additional ITA instruction during their first semester of teaching.

A more hands-off approach can be seen at University F. At this university there is a 4-hour orientation with general topics that all TAs are required to attend. There is also a summer TA training institute that is voluntary. The summer training includes 3 sessions: course design, planning activities and evaluation, and final course preparation. These sessions are followed by four 1-hour meetings with a faculty mentor ("STIA").

During the semester, TAs at University F can reserve a seat at monthly topical seminars. The university also has an Instructional Development organization with an information-rich website containing many pages of tips and online handbooks for TAs and ITAs. This website seems to be the primary source of information that is specifically designed for *international* TAs.

University C, described previously, takes a mentoring model in training ITAs; newly appointed TAs are referred to take part in consulting at the Center for Instructional Development and Research ("Consulting FAQ"). There they schedule ongoing consultations throughout the term. Their individual teaching consultant provides feedback on issues that require immediate attention and conducts end of the term teaching reports. ITAs also attend and participate in an annual university-wide TA conference.

At the UA, once ITAs have passed the language assessment, there is minimal mandatory training. There is one day of orientation to the university and a discussion of learner-centered pedagogy. Then all TAs must take and pass (with 95%) the TATO: TA Training Online. ITAs who do not pass the initial language screening, or who wish to improve their teaching ability, can take a teaching methods class for ITAs that includes simulated teaching as well as classroom observations. If ITAs feel their weakness is in pronunciation, they can take a pronunciation class through the Center for ESL, however, it is among non-matriculated students, and may be less rigorous than the ITAs desire.

There are also programs that vary by department. In the Writing Program, ITAs who are assigned to teach freshman composition must participate in a year-long preceptorship course alongside their domestic TA colleagues. Another program, Educational Psychology, has a weekly teaching seminar for all TAs and ITAs during all semesters of their teaching assignments, even when the TAs are in their 2nd or 3rd year. These examples show that some departments may be taking responsibility for the training of TAs and ITAs, but there is not centralized information about these policies and procedures. Furthermore, even with this pedagogical support, there still may be a need for training that is specific to the needs of ITAs, such as language practice or teaching strategies to scaffold learners around persistent pronunciation differences.

ITA research has grown since Bailey's groundbreaking publications on the "foreign TA problem" (1983, 1984). Most researchers have given training recommendations to improve various aspects of ITA, I next review some of those training suggestions .

Table 2.1

Summary of ITA Training Recommendations

Categories	Training recommendations	Source
Pedagogical & Cultural	Observe experienced TAs Train in lesson and assessment planning Use mid-term evaluations to gain feedback during the course of the semester	Numrich, 1993
Linguistic	Practice asking and answering questions	Numrich, 1993
	Teach communicative competence (cohesion & coherence strategies)	Hoekje & Williams, 1994
	Pronunciation & Vocabulary	Gallego, 1990
Interpersonal	'Attentiveness' (listening to students carefully, giving feedback on questions, deliberately demonstrating comprehension, creating a friendly atmosphere,)	Inglis, 1993
	Developing rapport with students, creating comfortable classroom atmosphere, heightening student interest crucial to success of ITA	Hendel et al, 1993
	Interactive personality and positive attitude correlate to high student evaluations of ITA effectiveness	Bailey, 1983

The initial reaction to calls for ITA training resulted in the creation of many linguistically oriented programs (Dick & Robinson, 1993, p. 8) since then, more focus has been directed to issues of cultural and communicative competence. Specifically, cultural and pedagogical recommendations include teaching US educational philosophies and assumptions, encouraging peer observations (of other experienced TAs or faculty mentors), explicit training in lesson/curriculum planning as well as implementing mid-semester student evaluations to get immediate feedback for making teaching modifications (Numrich, 1993).

If Linguistic skills are emphasized, Gallego (1990) observed that pronunciation and vocabulary miscues are the features most likely to cause communicative breakdowns. Numrich (1993) suggests that ITAs should practice asking and answering questions, and building this skill is also recommended later as a component of ITA 'attentiveness.' Hoekje & Williams (1994) focus on communicative competence and specifically recommend practicing strategies of cohesion and coherence to help ITAs signpost significant information during lectures.

Finally, several researchers place emphasis on interpersonal traits. In Bailey's (1983) summary of her dissertation research, she found that a positive attitude and interactive personality correlated to high evaluations of ITA effectiveness. *Interactiveness* is also cited by Hendel et al, (1993) who emphasize the importance of developing rapport with students, creating a comfortable classroom atmosphere, and heightening student interest as crucial to success of ITAs. Inglis (1993) highlights a similar quality, *attentiveness,* which is defined as listening to students carefully, giving feedback on questions, deliberately demonstrating comprehension, and creating a friendly atmosphere. These researchers found that the development of these interpersonal skills often allowed ITAs and international faculty to compensate for linguistic weaknesses.

The benefits of specialized training for international TAs extend beyond the ITAs' graduate participation, and their undergraduate instruction as ITAs. In the U.S., an increasing number of full time and adjunct faculty are non-native speakers of English (NNES), more than half of international graduate students remain in the U.S. after graduating, and many pursue jobs in academia (Gareis & Williams, 2004, p. 45). Once these NNES faculty are hired there is very little professional training that is specifically tailored to international faculty. In Gareis' 2004 survey, in the rare instance of training availability, it was conducted with either ITAs or with undergraduates, and for obvious reasons, faculty were unlikely to participate. Gareis also found reluctance to participate in such public programs when international faculty were nearing tenure or promotion for fear of negative repercussions of seeking assistance (p. 47). A mutually beneficial and preemptive solution is providing extensive training while ITAs are still graduate students.

In this section, I described several different approaches to ITA training. In a survey of research on training program components, Lewis (1997) found that the experience of teaching alone was not sufficient to improve student ratings (p. 9). Only with training, behavioral feedback, or consultations did ITAs demonstrate improved teaching achievement (p. 10). Additionally, as ITAs taught more advanced classes, their scores also improved (p. 10). This may be a result of having more accommodating, specialized students, or reflect that they had more years of experience as a TA. In the next section, I look at ways that various programs evaluate their ITAs' teaching ability.

2.1.4 Evaluating. The second goal of this dissertation is to investigate if the results of hiring tests, in this case the TBEST or TAST, has predictive criterion validity in predicting ITA effectiveness. For example, the SAT was designed to predict how well high school students would perform in college (Reynolds et al., 2006 p. 127). Similarly, I inquire if the scores on the TBEST or TAST predict how well ITAs perform as undergraduate instructors. In my quest to find the predictive utility of a hiring assessment, I must first find a valid effectiveness criterion on which to base my analysis.

Preliminary research shows mixed results when researchers have searched for the predictive utility of common language assessment tools in past studies. Jacobs and Friedman (1988) found correlation between screening tests and student evaluations of ITAs; whereas, Dunn and Constantinides (1991) found that the TOEFL scores were less useful in predicting ITA teaching success, but more accurate in predicting ITA lack of success. These results need to be tempered by research that finds that ITAs' scores may improve with additional experience (Kulik et al, 1985; Davis, 1991) and therefore the predictive validity may change as ITAs teach additional courses.

A summary of evaluation materials (Smith et al., 1992) finds that ITAs' effectiveness are commonly evaluated using:

- Student evaluations
- Surveys of ITA training program participants and staff

- Surveys of the ITAs' department heads or supervisors
- Evaluations of the amount of progress made by students of the ITAs as measured by course grades. (Smith et al., 1992, p. 3)

In addition to student and advisor evaluations, University G outlines an extensive procedure for gaining mid-semester feedback: fast feedback and structured mid-semester evaluation. The TA can administer the fast feedback forms at any point to identify areas for improvement in a quick and efficient manner ("Teaching," p. 11). The online handbook is very useful because full feedback forms are provided with alternate versions for lecturers, discussion courses, and lab instructors ("Teaching," pp. 13-19). There are also forms for fast early feedback, which cover student backgrounds and expectations, as well as later feedback, which cover more structural and pedagogical features of the class (p. 11).

The University G handbook also describes mid-term evaluation by faculty and includes sample evaluation forms ("Teaching," pp. 23-30). The mid-term evaluation has two components, concerning the TA and the course separately. Mid-semester evaluation by a supervisor is required, and it allows ITAs the opportunity to receive feedback while there is still time in the course to make modifications. The evaluations cover professionalism, teaching skills, attitude, aptitude, and an overall evaluation. The materials in this handbook are very instructive as they could be broadly applied in many different disciplines, and are structurally sound in providing both quantitative results that are easily statistically evaluated, and qualitative feedback which can give ITAs and advisors concrete suggestions for TA improvement.

There are two main choices for measuring ITA effectiveness at the UA, TCEs which are student evaluations, or advisor evaluations. The TCE is possibly flawed because of the perception that undergraduate students are biased against ITAs ("Understanding, " p. 3). However, interviewing departmental supervisors presents a conflict of interest, since they have hired the ITAs, they have a vested interest in affirming the ITAs' fitness-to-teach (Freeman, 1996). Since the TCE is more widely used and the scores are readily accessible, TCEs will be used as the criterion evidence in this study. Further explanation of this decision and of the TCE content is found in Chapter 3, section *3.6.3 Description of the TCE*.

All UA instructors administer TCEs at the end of each semester. The results are returned anonymously after grades are submitted. There are many variations of TCEs, the short form and the long form, as well as course-specific supplementary questionnaires created by individual instructors. However, there are four common questions that are on all versions of the TCE which cover overall impressions of teacher effectiveness, the amount students learn, course quality, and this instructor compared to others. These four questions will be the source of my criteria analysis.

Being an international graduate student and ITA is a difficult job. Not only do ITAs go through the normal academic challenges of graduate school, but they also may have cultural difficulties of unfamiliarity and homesickness and some additional language difficulties. Add to those challenges the incredibly demanding ITA occupation, which is exhaustively time-consuming and often done under great student criticism. While beneficial for paying for tuition and some personal expenses and providing teaching experience, international graduate students should not take application to an ITA position lightly, neither should hiring departments appoint ITAs flippantly. The variety of hiring, training, and evaluating standards for ITAs, highlighted in this review, point to the difficulties that universities face in matching their international student needs and institutional demands. It would be beneficial for universities that hire ITAs to have standard-setting guidelines to look to for guidance.

In this section, I discuss the assessment population in extensive detail. I look at the national and local enrollment trends, discuss the stereotypes that ITAs face in the classroom and survey ITA hiring, training, and evaluating procedures at universities across the United States as well as reviewing the suggestions from ITA researchers. This is a significant population to study, and the ramifications of effective ITA assessment effect not only the ITAs' funding and professional development, but also undergraduate education and institutional integrity. In the following section, I examine the process of developing a new assessment tool, TBEST, which may allow universities to streamline their ITA hiring process by providing a valid, reliable, and standardized method of assessing potential ITAs.

2.2 Test Functions

Before creating a new assessment tool, it is important to clarify the purpose of the exam. Tests can be primarily divided into norm-referenced and domain- or criterion-referenced testing. Then there are three basic functions that assessments fulfill, providing formative, interim, and summative feedback. Since each testing format gathers different information, for different audiences, and different applications, it is important to make the appropriate format decisions at the beginning of the development process. In this section, I define these assessment options and identify the appropriate variables of an assessment tool for this particular intended test application. Going through this decision-making process will assist others in the process of developing assessment tools for various other applications.

2.2.1 Norm-referenced assessment. Norm-referenced assessment frequently refers to mental-age or mental-level testing. This includes developmental assessment, measuring grade equivalents, and IQ tests. In a norm-referenced test, the sample is rated against the population norm, nationally, for example. Norm referenced tests measure "degree of attainment" (Anastasi & Urbina, 1997, p. 79) and without a "set of standard specimens" (p. 80) this isn't a possible method for evaluating ITAs' spoken English. Academic speaking is not a skill that has been (or could be) standardized, there is not even an agreed upon norm population for our sample of ITAs to be measured against.

2.2.2 Domain-referenced assessment. Domain- or Criterion-referenced testing emerged in the 1970's in the Education field as a method of testing content and performance objectives (Anastasi & Urbina, 1997, p. 76). A significant content domain is selected, and then broken down into smaller concepts, methodologies, or performance units. Then these units are tested to measure the degree of domain mastery. Anastasi and Urbina (1997) warn that identifying the test material is difficult and time-consuming, and without "careful specification and control of content, the results of domain-referenced testing could degenerate into an idiosyncratic and uninterpretable jumble" (p. 77). For this reason, the primary validity study in this dissertation examines the test *content* in the potential testing instruments, the TBEST and the TAST. All the standard setting and cut-score analysis will be worthless if the content is not appropriately matched to the assessment goal.

The use of domain-referenced testing also fits the intended application of the tests in this study. Domain-referenced tests are historically used "to check on pre-requisite skills, diagnose possible learning difficulties, and prescribe subsequent instructional procedures" (Nitko, 1989, in Anastasi & Urbina, 1997, p. 76). The tests described in this study are used to identify international graduate students who have the speaking ability to work as teaching assistants for undergraduate courses. Currently there is not an extensive university-wide training course at the UA to provide instructional follow-up for these students, but this dissertation makes recommendations for future development.

To further identify the appropriate test characteristics, I next examine the continuum of summative, formative, and interim assessment formats. To summarize the differences, these formats lie along the following spectrum: formative testing provides feedback to both the teacher and the student, and is used for instructional purposes. Interim testing, as the name implies, takes place in the middle of a larger instructional unit, and provides the student and the teacher feedback on the students' progress up to that point. Summative feedback, on the other hand, is only intended to provide information to a teacher or administrator for decision-making purposes, such as, if students pass a class, candidates meet a hiring requirement, or a teaching initiative merits ongoing funding.

2.2.3 Summative assessment. In the current education climate, formative assessment has been prioritized over summative assessment. However, as Taras found after publishing a much needed definitional article on the differences between Summative and Formative assessment (2005), many educators continue to have imprecise understanding of the distinction between the two forms (2008). In the absence of precise definitions, stereotypes seem to rule and 'formative assessment is [seen as] the antiseptic version of assessment and summative assessment has come to represent all the negative social aspects" (Scriven, 1967, p. 42). As Taras defines in her 2005 article, the process of all assessment "leads to summative assessment, that is, a judgment which encapsulates all the evidence up to a given point. This point is seen as a finality at the point of the judgment" (Taras, 2005, p. 468).

Summative assessments are also inherently political, as they are used to justify actions, whether hiring in the case of this study, or school and program funding as schools compete for increasingly limited state and federal funding. It may be these political and financial implications that make teachers untrusting of summative assessment. However, Taras reminds us that society naturally and rightly makes judgments, and "the misuse of judgments does not invalidate or minimize the necessity. It seems the very fear of the possible social misuse of assessment has distorted our view of it" (2008, p. 174). Rather, as summative assessments are designed, every effort must be made to identify the appropriate measurement and application of their results.

Hagstrom (2006) focuses on the accountability function of summative assessment, saying, "education experts commonly use summative assessment to certify the amount that individual students have learned and to provide an accountability measure for students and educational systems as a whole" (p. 24). She goes on to specify that summative assessment holds the "student accountable for what he [sic] has learned, and conversely holds instructors and programs accountable for what they have taught" (p. 24).

It is generally assumed that a wide gulf exits between summative and formative assessment, however, to paraphrase Taras (2005) and Scriven (1967): All assessment is initially summative assessment. If it stops there it is singularly summative assessment. If however, the teacher composes feedback, identifies what is missing, and what steps can be taken by the student to improve, then the summative assessment has become formative assessment.

An example of this transition can be seen in Vaden-Goad's (2009) study on increasing student motivation in a college math class. The periodic summative tests are repurposed as formative by giving students the option of replacing early low scores with higher scores on later exams (Vaden-Goad, 2009, p. 154). By definition, summative tests just provide the instructor with information, but when the tests are returned to the student and used to motivate or facilitate student learning, a hybrid test emerges, summative in design, formative in application.

2.2.4 Formative assessment. Formative assessment then, is basically summative assessment plus feedback. More specifically, "for an assessment to be formative, it requires feedback which indicates the existence of a 'gap' between the actual level of the work being assessed and the required standard. It also requires an indication of how the work can be improved to reach the required standard" (Taras, 2005, p. 468). Ramaprasad (1983) expands this definition by clarifying that feedback is "information about the gap between the actual level and the reference level…which is used to alter the gap in some way" (p. 4).

Black goes a step further and defines the possible uses of formative tests:

1. classroom dialog
2. peer & self-assessment
3. comment-only marking, dialog in writing
4. formative use of summative tests, using test answers as an occasion for formative feedback. (Black, 2009, p. 519)

Black hastens to qualify that the benefits of formative assessment are strongly contingent on teachers modifying their lessons based on the results of the formative feedback. Black notes that class modifications are usually "limited to one-to-one Teacher-Student interaction, but it would be beneficial for this to have whole-class dialog before the Teacher decides how to intervene" (p. 521).

Hagstrom, a speech pathologist, sees formative assessment as a component of constructivist learning, in the tradition of Piaget and Vygotsky. She sees formative assessment as a way to integrate assessment into the wider contextualized educational experience of students (2006, p. 26). Similarly, Ecclestone and Pryor (2003) argue that: "formative assessment is better conceived of as an interactive pedagogy based on constructivist ideas about learning and integrated into a wide range of learning and support activities" (p. 472).

Taras (2005) further describes the relationship between formative assessment and summative assessment: formative assessment encompasses and justifies summative assessment, clarifying how the parameters have been addressed, and what needs to be done next (p. 470). For these two reasons, she maintains that formative assessment is more important than summative assessment for most contexts.

2.2.5 Interim assessment. Interim assessment is an emerging assessment niche between "state-level, once-a-year summative tests and day-to-day formative assessment used as part of classroom instruction" (Shepard, 2009, p. 35). However, it is still regarded to be in the conceptual stage; similar to mini-summative tests, and not something that limited assessment funds should be used on (p. 35). The proponents of interim assessment, Perie et al. (2009), identify instructive, evaluative, and predictive criteria that can be used to identify when interim assessment may be beneficial (p. 35). However, as Shepard notes in her reply, those criteria are not met. Specifically, the instructional value is said to be no greater than test banks and end of chapter test that are included in most text books, and the predictive value of identifying at-risk students is not shown to be any more accurate than what an observant teacher can identify (p. 36).

After defining and discussing these two types and three purposes of assessment, it is clear that the assessment needs of this study call for summative, domain-referenced assessment. The graduate college needs to screen potential ITAs prior to hiring. The assessment is not a part of a larger instructional course, nor is individual feedback given to the examinees. The primary purpose is to test for the pre-requisite speaking skills for becoming an ITA. However, as previously stated in this section, while the primary purpose of this assessment goal is summative in nature, there is always the possibility for it to be further developed to give formative feedback and guide ITA training.

2.3 Test Formats

There are four main assessment formats, or constructs, for measuring spoken English ability according to a survey of tests currently used by universities to evaluate non-native English speaking (NNES) graduate students: single-skill assessment, oral proficiency interviews (OPIs), integrated-skills assessment, and task-based assessment. There are other constructs for assessing spoken language as well, such as self assessment, information gap, picture stories, giving directions, retelling stories, and translating (Underhill, 1987), but these forms of assessment are better suited to formative assessment in classroom settings.

2.3.1 Single-skill assessment. The single greatest complaint against ITAs is pronunciation. Virtually every article on this subject seems to include student anecdotes of communicative misunderstandings. It is unsurprising then, that a single test of pronunciation would be used in some cases to screen ITAs. A single-skill assessment, a test of word-level intelligibility, is described by Isaacs (2008) to screen incoming ITAs. Isaacs notes that while "pronunciation is only one of a range of factors that can affect ITAs' ability to carry out their instructional duties, poor pronunciation is the most overt problem associated with ITAs" (p. 560). In this exam, taped speech samples (based on TSE prompts) are rated according to how well the rater understands every word that was spoken.

There is an obvious detriment to using such a narrowly focused screening assessment tool. Testing pronunciation in isolation, cannot possibly predict proficiency in all the elements of spoken communication that are required to accomplish the tasks that ITAs must perform. This testing construct takes the complex Communicative Competence construct in the opposite, and equally impractical, direction. Furthermore, it was easy for ITAs to master the test instead of mastering the overarching English ability. Apparently, ITAs soon learned to speak painfully slow and to over-enunciate every word and therefore gain a high score. In fact, the speaker with the highest pronunciation score in the study had speech so slow that it was "hard to link the parts of the sentence together" (Isaacs, 2008, p. 570).

Isaacs concludes that the intelligibility test is a valid test for ITAs, considering the complaints about pronunciation that undergraduates usually make. However, he also recognizes that this exam needs to be part of a battery of screening tools that are used to evaluate ITAs for teaching positions. I maintain that this method is too flawed to be useful, even within a larger battery of tests: the scope is too narrow, it is inauthentic, and easily fooled.

2.3.2 Oral proficiency interviews (OPI). The OPI may take place formally, according to ACTFL structure and standards, in the SPEAK version of the TOEFL, or casually as an actual conversation, and may include a role-play component. For many people this is the "best and fairest measure of oral ability" (van Lier, 1989, p. 490; Stevenson, 1983; Lowe, 1986). However, the task of administering the OPI may seem "insurmountable" to larger populations (Gonzalez, 1988). The assessment tools that I analyze (the TBEST and the TAST) are both used for large-scale examinee placement while the OPI may be more useful for giving educational, formative feedback to a smaller population.

There are two main objections to the OPI being used as a placement or selection tool: reliability in test delivery and reliability in rating. Brown (2003) highlights the weakness of the OPI in that while its form mimics the "dynamic nature" of natural conversation, there is the potential for great variation in every interview. The variations among interviewer strategies can lead to very different test performances. Over time, interviewer differences may become magnified, creating different levels of challenge (Brown, 2003, p. 19). Some candidates may receive more feedback (follow-up questions, questions re-stated, or question sequencing that is more 'leading' than others) and in the end, raters evaluate very different language samples, which in turn, result in scores that cannot be comparably evaluated, and testers are not given equal speaking opportunities. Occasionally, role-play is included in an OPI. This method is also unreliable; in role-play situations the examiners rarely interact in the exact way with all examinees, so some ITAs would have very different role-play scenarios, and raters tend to tire quickly and lose objectivity (Plakans & Abraham, 1990, p. 78).

Even if OPI administrators do follow a standardized script when interviewing candidates, the rating is still very variable. In van Lier's (1989) support of using OPIs, he recommends "sidestepping the issue of construct validity altogether and be[ing] satisfied with measuring whatever oral language use happens to be elicited by the OPI since it is the best instrument available" (p. 501). He acknowledges that conversation is difficult to rate, both in knowing “what and how” to rate, but also because the interviewer would be busy participating in the conversation. Rater reliability among a large candidate pool is difficult to establish when all raters may be evaluating very different language samples and based on holistic impressions, rather than the constancy of having standardized prompts and an analytic rubric as in the case of the TBEST and TAST.

The TBEST and TAST are both designed to test a large number of candidates quickly and to place candidates along a standardized scale for selection or placement purposes. On the other hand, an OPI is commonly administered to assess communicative competence of a smaller population, with local results (ie: for either teacher or student benefit, within a classroom) and to give formative feedback. The utility and function of the OPI makes comparison with computer-based, standardized assessment inappropriate.

2.3.3 Integrated-skills assessment. The revised version of the TOEFL iBT is an example of an integrated skills test. On the speaking portion of the iBT, the TAST (the TOEFL Academic Speaking Test), heterogeneous tasks are used to measure a single construct, speaking proficiency (Lee, 2006, p. 132). The tasks assess listening-speaking, reading-speaking, and stand-alone speaking ability. Lewkowicz (1997) argues for integrated task testing because:

- Test-takers are less likely to be disadvantaged due to lack of information on which to base their argument (Read, 1990; Weir, 1993) and
- Validity is enhanced by simulating real-life communication tasks in academic

> contexts (Wesche, 1987). (Lee, 2006, p. 134)

However, as Lee goes on report, there are concerns about the dependability of integrated skills assessment. Lee's study uses generalizability theory (G- theory) to make simultaneous measurement of more than one facet; studying the interaction of task types, number of tasks, and raters. Answering the question of task construct validity, Lee found that there was a difference in the scores on the integrated and independent speaking portion of the test, but not large enough to justify not combining all scores as a single composite score (p. 160). This result validates the TAST for integrated assessment purposes, such as for admission for international graduate students, or as an exit examination from an English remediation program. However, integrated assessment may not be appropriate for high-stakes speaking assessment as in the case of ITA hiring assessment tools.

There are also arguments against integrated skills assessment in this context. Since the primary goal of ITA assessment is measuring academic spoken English ability as pre-requisite for teaching, it is important that the skill is *highlighted* and examinees are not incidentally evaluated on timed reading or isolated listening ability at the same time. Timed reading is a very unusual task for ITAs (or anyone, for that matter) to have to perform, and on the TAST, examinees must read a 75-100 word passage before speaking (as opposed to 10-25 words in the TBEST). In real-life situations, ITAs have extended time and additional resources (using dictionaries or asking a friend) to assist with reading comprehension, if need be. Also, listening is only a small portion of understanding oral input; but on the TAST, examinees must respond to a 60-80 second listening excerpt. In a real communicative event, the ITA would have non-verbal cues, as well as the opportunity to ask for clarification or re-statement, to assist in comprehension. The only verbal input on the TBEST is in a video clip, so examinees have non-verbal cues to augment their listening ability. Since these input realities are not controlled in the integrated assessment on the TAST, and because the test content focuses on very general English speaking tasks, I do not think the TAST is the most valid test for assessing incoming ITAs. It is widely used though, and more valid and reliable than single skill assessment or OPIs, and therefore it will be used for comparative analysis in this study.

In 2006 the Educational Testing Service (ETS) conducted a standard setting session to set cut scores that would allow the TAST to be used for screening ITAs, and not just as an academic admittance tool. The team included a diverse representation of professionals who work with ITAs in various capacities from 18 universities. In this study, the researchers asked panelists to set cut scores on the TAST through the benchmark method (Faggen, 1994). Panelists reviewed the TAST's rubric criteria and selected their optimal score point (1, 2, 3, or 4) at each criterion. The combined optimal score then represented the lowest acceptable response that would indicated an ITA candidate was minimally acceptable for hiring (Wylie & Tannenbaum, 2006, pp. 5-6). The standards were first set individually, then results were discussed, and then in a second round of individual ratings, which were averaged to express the panel's consensus.

In the second stage of the standard-setting, the panel tried to establish a correlation between a TSE score of 50 (an often used benchmark for ITA hiring) and the TAST. In this stage panelists were simply asked: "Given the description of what a candidate with a score of 50 on the TSE could do, how would that candidate perform on the TAST?" (Wiley & Tannenbaum, 2006, p. 8). This seems like a questionable method given that panelists were considering a hypothetical candidate, taking remarkably different tests, with the TSE focusing on a wide-range of general tasks, and the TAST being a more intensive integrated-skills speaking assessment (p. 8). The researchers concede that this judgmental standard-setting method was selected, instead of an empirical approach to save time, and not on the basis of validity. The panel decided that the minimum TAST score for ITAs should be 23, though the TAST 26 was the likely equivalent to the TSE 50 (Wiley & Tannenbaum, 2006, p. 11). In my comparative analysis of the TBEST and TAST, I focus more on the content validity and criterion-validity (predictive utility) and not on the particular cut scores of each.

2.3.4 Task-based assessment. Task-based assessment allows the examiner to assesses language ability regarding specific content. An example of this is the GSLPA (Graduate Students' Language Proficiency Assessment) in Hong Kong, which was created in both English and Chinese to predict students' success in business communication (Lumley & O'Sullivan, 2005). The GSLPA evaluates testers' ability to give presentations, leave work-place related phone messages and have work-related interviews (p. 417). Similar to the need demonstrated in Hong Kong for business English assessment, many university administrators in the United States are calling for a test to be developed which specifically measures English speaking ability of potential ITAs.

***2.3.4.1 Theoretical basis for T-BEST*.** The TBEST, originally developed in 2007 by Dr. Jun Liu and operated by Mind Works in China for oral proficiency testing of middle and high school students represents an efficient system for large-scale spoken language testing.[4] With a successful track record of assessing tens of thousands of students in China, it seemed adaptable to its current application: large-scale English proficiency testing for international teaching assistants (ITAs) in the American university setting.

Tasks and rubrics existed from the original rounds of testing a variety of student populations and proficiency levels in China; these were studied and refined for use with ITAs rather than it being necessary to completely create a new test. In China the tasks are imitative, communicative, descriptive, comprehensive, and argumentative to suit the assessment context for beginning to intermediate level English language learners. These tasks were modified to cover descriptive, interpretive, argumentative, and comprehensive classroom management tasks for the TBEST that is used to assess ITAs. Further explanation of the TBEST assessment content is found in section *3.6.1 Materials: TBEST* and domain analysis of the TBEST is in section *4.2.1 Comparative domain analysis.*

In domain-referenced assessment, the assessment domain must be clearly defined for the test to be valid. Anastasi & Urbina outline the process of identifying the test domain: “the content domain to be sampled must be widely recognized as important. The selected domain must then be subdivided into small units defined in

[4] In China, TBEST stands for Task-Based English Standards Test. At the University of Arizona, the modified TBEST stands for Task-Based English *Speaking* Test. For a description of the Chinese TBEST see http://tbest.eassol.com/index.html.

performance terms" (Anastasi & Urbina, 1997, p. 77). In designing the TBEST, the ITA domain was defined through a consulting process with department heads, ITA advisors, and graduate TAs, deciding on the most common speaking tasks in the TA setting. These members all view the ITA position from different perspectives so a complete picture could be assembled. Since there is a lack of consensus among scholars and researchers in second/foreign language learning and teaching regarding the meaning of a task, I offer ten representative interpretations below before I define what a task means in T-BEST.

2.3.4.2 Definition of a task. While there are differences of opinions about the definitions of a *task* in language assessment, three features are commonly present: it is an activity with a definite objective, based in the real world, and with a focus on meaning over form. Several researchers specify the classroom context, with a specific task objective relating to cooperation between learners manipulating and processing the target language (Crookes, 1986; Prabhu, 1987; Nunan, 1989; Lee, 2000; Bygate, 2001; Ellis, 2003). Secondly, there are an infinite number of activities that can be reflected in tasks. As Long (1985) stated, tasks are the everyday things that people do in work, play, and everywhere in between: making reservations, writing checks, weighing a patient, finding a destination, and so on. It is important though, that the emphasis is on real-world language needs (Skehan, 1996; Ellis, 2003). Thirdly, if the teacher has effectively designed the task, learners do not focus on form, but are engaged in the process and communicate freely. The focus of assessment then is on meaning, on the objective being met (Nunan, 1989; Skehan, 1996; Lee, 2000; Bygate, 2001; Ellis, 2003). A summary of these definitions is in Table 2.2.

Table 2.2

Summary of Task Definitions

Year	Author	A task is/are…
1985	Long	The everyday things people do in work, play, and in between.
1985	Richards, Platt & Weber	An activity or action carried out as the result of processing or understanding language.
1986	Crookes	A piece of work with a specified objective for education, work, or to elicit research data.
1987	Prabhu	An activity that requires learners to arrive at an outcome through a process of thought that is regulated by teachers.
1989	Breen	A brief practice exercise or a complex plan that requires spontaneous communication of meaning.
1989	Nunan	A classroom activity in which learners comprehend, manipulate, produce, or interact in the language, while principally focusing on meaning rather than form.
1996	Skehan	An activity in which meaning is primary, there is some connection to the real world, task completion has priority, and assessment is based on outcome.
2000	Lee	An activity in which the objective is obtainable only by interaction among participants, there is a focus on meaning, and learners must manipulate or produce the target language as they perform the activity.
2001	Bygate	An activity which requires learners to use language, with emphasis on meaning, to attain an objective.
2003	Ellis	A work plan; A task involves a primary focus on meaning; A task involves real-word processes of language use; A task can involve any of the four language skills; A task engages cognitive processes; and a task has a clearly defined communicative outcome.

According to Lewis' 1997 survey, Typical TA Tasks include:

Attending class, taking attendance, holding office hours, writing exam questions, grading exams, proctoring exams, conducting review sessions, grading homework, grading papers, maintaining class records/grades, making overhead transparencies and handouts, teaching a class (p. 2).

This list is representative of what Lewis defines as the original purpose of the TA position. Originally, the TA position was based on an apprenticeship model, assisting a professor in a specific course, but seldom having direct student contact (Lewis, 1997, p. 1). The university boom that began in the 1960's though required the university to make structural changes, which included utilizing TAs for more instructional purposes (p. 1). The responsibilities of TAs continues to expand, with many TAs having more direct student contact and limited professor oversight.

At the UA, the primary job skills of ITAs, as determined by the ITA assessment panel that I participated in, are:

- Teaching Content
- Explaining discipline-specific theories
- Explaining concepts, homework, and various other elements of the class
- Justifying grades (in office hours)
- Classroom/Lab management

Following Anastasi & Urbina's method of creating domain-referenced assessment content, having selected the significant content domain, that domain was broken into smaller performance units (Anastasi & Urbina, 1997, p. 76). Then these units: a Decriptive task, Interpretive task, Argumentative task, and the Comprehensive video task are tested to measure the degree of domain mastery.

In the TBEST, a task is defined as a communicative performance in the absence of the interlocutor where the examinee is required to complete a specific function. The description of these functional units are as follow: describing an incident (descriptive task), describing/paraphrasing the meaning of an ambiguous statement (interpretive task), expressing an opinion or taking a stance (argumentative task), and interpreting and evaluating a classroom scene in a video clip (comprehensive video task). It is believed that the combination of these four tasks reflect the overall English speaking proficiency of the examinee in a testing situation specifically tailored to assess the speaking skills required in an ITA position when measured holistically against the ten-point rating scale.

2.4 Conclusion

In this chapter, I cover the three areas of study which could require additional background knowledge to supplement the later analyses of this dissertation: ITAs, Test Function, and Test Format. I begin by reviewing the relevant literature describing ITAs, stereotypes about ITAs and the hiring, training, and evaluating procedures that are commonly implemented at various universities in the United States.

Having established that information about the population of this study, I focus on the specifics of test development. I first summarize the different functions that assessment can perform, beginning with norm- and domain-referenced testing. Then I look at three types of assessment applications: summative, formative, and interim. Finally, I examine the possible test formats, or constructs, that could potentially be used for assessing spoken English of ITAs. I eliminate single skill and OPIs as viable hiring assessment tools, and instead focus on integrated skills, as in the TAST, and task-based assessment, as in the TBEST. Since the TBEST is a newly developed assessment tool, I summarize literature that supports task-based assessment and the definition of a task.

With this literature foundation set, I proceed to a description of my research methods and the materials used to collect data.

CHAPTER 3

METHODS AND MATERIALS

This chapter presents an overview of the methods and materials that inform this dissertation. This overview is intended to orient the reader to my guiding research paradigm and describe the roles of the researcher and participants. Next, I summarize the analyses procedures that I employ, document and content analysis, statistical analysis, and qualitative contextualization. Finally, I provide a detailed description of the materials that I use for data collection, two tests, the TBEST and TAST, as well as student evaluations, and an ITA survey. As stated in Chapter 2, the TBEST is an English language assessment developed in China[5] that has been adapted for use for assessing ITAs at the University of Arizona, and this dissertation contains the first formal description of this modified assessment tool.

3.1 Mixed-Methods Paradigm

To respect the complexity and high-stakes implications of assessing ITAs, this dissertation follows a mixed-methods paradigm. As Patton (2002) states, the term *triangulation* is used to metaphorically represent the multi-faceted course of inquiry common in qualitative research (p. 247). The logic is that any method has weaknesses and strengths, and by utilizing a purposeful mix of methods, I approach my research with "an arsenal of methods that have non-overlapping weaknesses, in addition to their complementary strengths" (Brewer & Hunter, 1989, 17).

[5] TBEST was coined by Dr. Jun Liu (2007) article, and copyrighted by Mindworks in China. (Liu, J. (Ed.) (2008). *Teaching English in China.*

Denzin (1978) identifies four types of triangulation: data, investigator, theory, and methodological. In this dissertation, I employ both data triangulation and methodological triangulation. As I outline in section 3.4 *Materials,* the data that I analyze come from two different standardized tests (TBEST and TAST), student evaluations (TCEs), and a survey of ITAs . By including data from mixed sources, I am able to lessen the impact of bias from the sources, and find overall trends.

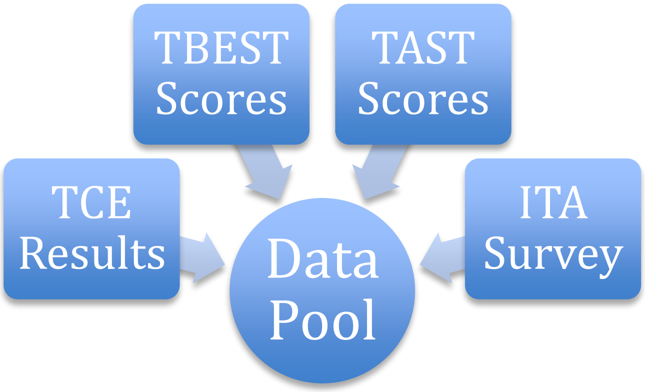


Figure 3.1. *Data Triangulation Elements*

The analyses methods that I use, outlined in 3.5 *Analyses Procedures,* include document and domain analysis, statistical analysis, and fixed response survey. This diversity of analyses types reveals the significance of my data. For example, trends in narrative data can be substantiated or challenged through statistical analysis. This deliberate process is necessary because "each method reveals different aspects of empirical reality, multiple methods of observations must be employed" (Denzin, 1978, p. 28).

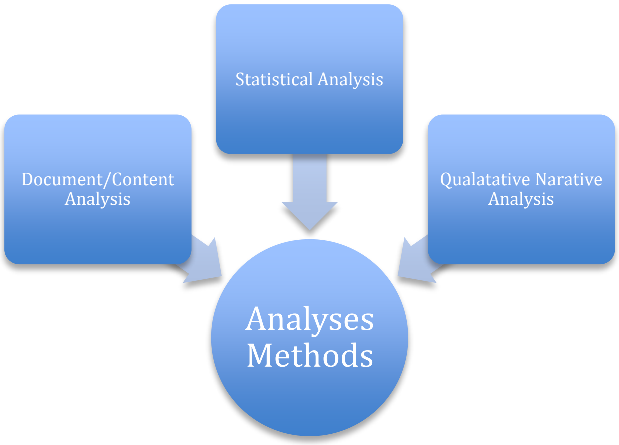


Figure 3.2. *Methodological Triangulation Elements*

3.2 Role of the Researcher

As the researcher, my role has varied over the course of the study. All participants initially met me when they came to take the TBEST; I was the test administrator and explained the test content and procedure before they began the exam. If participants consented online, they knew their TBEST scores were being used in this study and this was an opportunity for them to ask questions. My role explicitly excluded test rating and score reporting. My second contact with participants came after one or two semesters passed, during which time qualified ITAs taught their undergraduate courses. I contacted ITAs by email and requested that they participate in a survey about their TA experience.

3.3 Participants

Every academic year, the University of Arizona hires approximately 3000 Graduate TAs. Of those, around 900 are International TAs, and approximately 200 of those ITAs are new to the UA each year. The Arizona Board of Regents (ABOR) requires all ITAs to take and pass an English speaking assessment before being hired as a TA. Many of these graduate students have taken the TOEFL iBT , which includes the TAST, before arriving, as this had previously served a dual-purpose as a graduate school admittance requirement and as an employment screening tool. All TAST scores that were reported by students taking the TBEST were used as comparative data. This amounts to 58% (n= 193) of ITAs who were tested using the TBEST. From this pool of examinees I searched for those ITAs who had taught as the primary instructor and who had been evaluated by the TCE for the predictive criterion portion of my analysis. Unfortunately, only 11 ITAs met all of these requirements. I report my findings, but also recognize the limitations of this small population. Finally, I conducted a follow-up survey of all 335 students who took the TBEST, and 119 examinees chose to participate. The survey was anonymous, online, and covered the issues of their hiring process, the training they received or would like to receive, and the ways their teaching was evaluated as well as a question about their overall satisfaction.

Over 300 ITAs were tested using the TBEST assessment tool during three semesters from Fall 2008 through Fall 2009. The ITAs represent great diversity in both country of origin and academic departments. The ITAs in this study come from 30 countries, 60% of them come from China or India, and the remaining 40% come from various countries in Europe, the East and Middle East, and the Americas. These graduate students also represent over 50 different academic departments, almost evenly split between the Humanities and the Sciences.

3.4 Ethical Considerations

Since this study utilizes potentially sensitive information and opinions, no personal indentifying information is reported. Identification numbers were randomly assigned for test score reporting, and surveys were conducted anonymously. No compensation was given for participation in this study. ITAs participated simply for the opportunity to help improve the hiring and working conditions for other ITAs at the university.

3.5 Analyses Procedures

The analyses procedures that I conduct are domain analysis, predictive criterion analysis, and qualitative contextualization.

3.5.1 Parallel test forms. Before I analyzed evidence for content validity for the TBEST and TAST it was vital that I confirm that they are not parallel forms of assessment. To do this I correlated TBEST and TAST scores for those ITA examinees who have taken both tests using the Pearson correlation. Once I affirmed that these two assessment tools are assessing reasonably different skills; I proceeded with my domain and criterion analyses.

3.5.2 Domain analysis. Evidence of domain validity is the first goal of this study. As Reynolds (1997) stated, domain-based validity evidence is preferred for establishing the validity of academic achievement and on tests used in the selection and classification of employees (p. 125) so this is a justified route of inquiry for examining an ITA hiring prerequisite. In order to identify skills in the ITA domain, I first looked at an explanation of TA responsibilities at the University of Arizona, as determined by a panel of experts and stakeholders in the ITA hiring process.

At the UA, the primary job skills of ITAs, as determined by the ITA assessment panel, are:

- Teaching Content
- Explaining discipline-specific theories
- Explaining homework answers
- Justifying grades (in office hours)
- Classroom/Lab management

Next, this list of ITA job skills was compared to a self-report from ITAs who have passed TBEST and subsequently taught for at least one semester. The test content on both the TBEST and the TAST were then compared to the most prevalent skills reported for ITAs. Domain validity is important to establish because tests are only valid to the extent that their results are used for their intended purposes. This document analysis should measure which test is better suited for this assessment purpose.

3.5.3 Predictive criterion analysis. The second goal of this study is to evaluate the evidence of predictive criterion validity for both the TBEST and the TAST. Concurrent criteria are assessments that are administered at the same time and Predictive criterion compares preliminary scores with a measurement taken after the passage of time (Reynolds, 2006, p. 127).

To measure the evidence of predictive assessment validity for both the TBEST and the TAST, I analyze Pearson's correlations of the TBEST and the TAST to TCE scores, on four questions that assess different facets of teaching effectiveness, to find which test has a higher correlation and is therefore more predictive of teaching effectiveness.

The results from each of the four questions are reverse scaled to convert the result to a single score. Those numbers are then correlated to the ITAs' TBEST and TAST scores to analyze predictive validity of both the TBEST and TAST.

Table 3.1

Demonstration of TCE Scaled Results Conversion

Respondent %	Likert Scale	Raw score
20	A = 5	100
60	B = 4	240
10	C = 3	30
10	D = 2	20
0	E = 1	0
100%		390/100= **3.9**

There were two main options for measuring ITA effectiveness, TCEs which are student evaluations, or advisor evaluations. The TCE is possibly flawed because of the variable features listed in section 3.4.4, but those factors are acknowledged and accounted for. However, interviewing by departmental supervisors also presents a conflict of interest, since they have hired the ITAs, they have a vested interest in affirming the ITAs' fitness-to-teach (Freeman, 1996). Therefore, TCE scores are used, anticipating a lower score mean than for native TAs or professors, evaluating the ranking of ITAs within the group of other ITAs only. To account for potential variance between different content classes, the TCE reports are reported by discipline.

Another reason for choosing the TCE instead of advisor reports as the criteria data is criteria contamination. Criterion contamination occurs when scores on the predictor and the criterion are not independent. For example, if instructors are aware of students' performance on an aptitude test, that might influence their evaluation of the students' performance in the class (Reynolds, 2006, p. 128). In a similar way, since the ITA advisors know the hiring cut scores, and what the ITAs scored in order to be hired to teach, this may influence their evaluation of the ITAs' teaching ability. On the other hand, students are not informed of the assessment and hiring process that ITAs go through, much less what their particular ITA scored, therefore their TCE scores are not influenced by the predictor score. Further validating student evaluation, in the OIRE report, students (.85) have higher inter-rater reliability than peer rating (.57), so student input should not be overlooked ("TCE Guide," p. 15).

A final reason is based on the results of my follow-up survey. I asked ITAs who had taken the TBEST how they would like their teaching to be evaluated. When given multiple options for evaluation, the most desirable methods of evaluating ITAs' teaching are: end-of-semester student evaluations (4.07 out of 5), student mid-term evaluations (3.97), and advisor observation (3.93). Since the current evaluation method (TCEs) is the also the most desirable method of evaluation according to the ITAs in this study (n=100), it is reasonable to use the TCE scores to measure ITA effectiveness. Currently, mid-term evaluations and advisor evaluations are not systematically conducted, but that can be an option for future program development.

3.5.4 Qualitative contextualization. Following my discussions of assessment domain validity and predictive criterion analysis of ITA effectiveness correlations, it is important to contextualize the results within a descriptive framework of the UA's ITA population. I also collected qualitative data in order to make training and evaluative assessment recommendations based ITA feedback.

I collected demographic data from ITAs and from hiring records about departments and countries of origin. I also use a follow-up survey to poll ITA attitudes about ITA training and evaluation. The surveys are conducted anonymously online. I analyzed the survey results by examining the rankings that ITAs gave different training and assessment options, and considered the patterns between answers in the Hiring, Training, and Evaluation themed-questions and the correlations between fixed-response questions and the short answer question.

3.6 Materials

Two assessment tools are compared throughout this study, the TBEST and the TAST. I also utilize TCEs (end-of-semester student evaluations), and an ITA follow-up survey.

3.6.1 Description of the TBEST. The Task-Based English Speaking Test (TBEST) is a testing platform that was initially used to evaluate middle- and high-school students' English in China[6], and which was adapted for use as an ITA assessment tool in the United States at the University of Arizona. The original testing platform was a computer-based test, with 4 task types: Imitative, Communicative, Descriptive, and a Comprehensive/Combined task. In repurposing the TBEST, a committee of experts and stakeholders was assembled with representatives from the primary hiring departments

[6] In its original form in China, TBEST stands for "Task-Based English Standard Test." http://tbest.eassol.com/index.html

(the Graduate College, English, Math, Chemistry, and Engineering) and testing experts (the Second Language Acquisition and Teaching doctoral program, the University Learning Center and the Center for ESL). This re-development team included participants representing the needs of administration, department heads, teaching advisors, and graduate TAs. These experts were consulted about ITA skill requirements to create task types and test questions specifically tailored for the ITA assessment context.

The panel summarized ITA duties as: teaching content or leading lab sessions, explaining discipline-specific theories, explaining homework answers, justifying grades, and classroom management. As a panel, we agreed that two of the existing task types in the TBEST framework, Descriptive and Comprehensive tasks, could be used to measure ITA duties and added the Interpretive and Argumentative tasks-both of which require higher language ability than would have been appropriate for the original Chinese student assessment population.

The Descriptive task is used to assess ITA ability to teach basic content. The Interpretive task, interpreting a proverb, is similar to being able to explain a theory, especially since many theories utilize metaphors, like ‘scaffolding’ in Second Language Acquisition theory. The Argumentative task tests the ability to justify a debatable position, as in justifying grades or homework answers. The most unique aspect is the Comprehensive task, which on the UA TBEST, is a video of an unsuccessful teaching scenario. When examinees respond to the video it allows raters to assess their ability to interpret inter-personal behavior in a classroom and to justify opinions about the teacher and/or students' actions. A summary of the TBEST tasks is presented in Table 3.2.

Table 3.2

Summary of TBEST Test Content

Prep/Speaking time	**TBEST task**	**Addresses this TA task**
20s/60s	Descriptive	Teaching content
20s/60s	Interpretive	Explaining Discipline-specific theories and homework answers
20s/60s	Argumentative	Justifying grades
2min view video, 20s/90s	Comprehensive: Classroom scenario video	Demonstrates awareness of complex classroom management issues

ITAs taking the TBEST spend approximately 30 minutes in the test facility; administrators spend 10-15 minutes explaining the test procedures and familiarizing students with the test instructions that they will receive, and then the test itself takes 10-12 minutes. All task prompts are 10-25 words long.

A sample descriptive task may require the examinee to describe what animal would make the best pet. An interpretive task would be to interpret the meaning of an ambiguous saying or proverb like "You can never step in the same river twice." An argumentative question might ask: "Which is better, to spend your money when you earn it or to save it for the future?" Sample teaching scenarios in the videos include a teacher who writes messily on the board and is oblivious to student requests to write legibly. There are no TBEST preparation materials other than the suggestion to practice speaking on the phone to native English speakers to simulate communicating without non-verbal cues during TBEST. This means that TBEST speaking samples are closer to examinees' actual speaking ability than the potentially overly-prepared answers on the TAST.

The score means for ITAs in this study is 7 (Table 3.4); true beginners' scores would fall in the 1-4 range, and 8-10 scores generally define expert speakers, with 10 describing *academic* speech of expert[7] English speakers. The examinee population is self-limited through the academic screening that takes place in the graduate school admissions process. Since these ITAs are pursuing advanced degrees in their discipline, in a second language (English), their language ability is obviously higher than the casual foreign language student.

As with the TAST, each language task is rated by two trained raters and averaged. If the scores differ by more than 1 point, then the sample is re-rated by a third rater. Each task is rated on a scale from 0-10 in the following four categories: Overall Performance, Accuracy (in word choice and grammar), Fluency (in pronunciation and intonation), and Organization (at macro- and micro- level). These categories are evenly weighted in calculating the examinee's score (See Appendix A).

In developing this rubric, we (the UA TBEST development team) began with the Chinese TBEST rubric as a guide and made adjustments according to our distinct test population and test content. We also met with our raters after one round of testing to identify categories that were confusing, or wording that could be conflated. For example,

[7] The term *expert* in this case refers to both native speakers of English and those who, though not native English speakers, have lived in an English-speaking environment and developed very high communicative competence in English. *Expert* does not have the same connotation of setting unattainable competency standards for second language learners as does native speaker.

the distinction between Accuracy and Fluency was not initially defined in a manner that was consistent or easy for the raters to judge. As a remedy, we re-worded the categories to clarify for the purposes of this test, that Accuracy pertains to grammar and word choice and Fluency is pronunciation and intonation. We also went through the various rating levels to make sure that the skills were represented in a consistent gradient in the low, mid, and high ranges.

In addition, we conducted two rounds of initial rater training. Our first round of testing revealed some tendency toward a halo effect-there was unusually high inter-criterion correlation that lead me to believe that our raters were rating holistically instead of rating each criterion (overall performance, fluency, accuracy, organization) distinctly. That trend is no longer present. After that training, we have conducted annual re-norming sessions.

3.6.2 Description of the TAST. The Test of English as a Foreign Language, Internet-Based Test (TOEFL iBT), was the previous ITA screening tool at the University of Arizona, and is well known by most educators. The iBT contains multi-skill assessment, but for the purpose of this study I reference only to the Speaking section results. I refer to this test as simply the TAST (the TOEFL Academic Speaking Test) to distinguish it from other TOEFL exams, such as the TSE (Test of Spoken English) and SPEAK (Speaking Proficiency English Assessment Kit) tests.

The complete iBT takes 4.5 hours with the TAST lasting 20 minutes. The TAST contains 6 tasks: 2 descriptive & argumentative tasks, 2 tasks based on summarizing reading & listening material, and 2 tasks based on listening input such as a lecture or conversation. Sample topics in the first section include describing a vacation or justifying a preference for living in the city or country. The second section could ask participants to summarize input about airline fee increases based on a written announcement from the company president and listening to a conversation between passengers. In the final activity, participants may perform a task like listening to a lecture on voicemail etiquette and then leave a voicemail that follows a specific prompt like "leave a voicemail to change an appointment with your dentist" or "leave a voicemail to congratulate a colleague on promotion" ("TOEFL Speaking"). Listening passages are 60-90 seconds long, and reading passages are 75-100 words long ("TOEFL iBT Tips"). A summary of TAST times and tasks is in Table 3.3.

To help candidates prepare, there are extensive TOEFL preparation courses offered at ESL centers, online, and through self-study. The availability of training courses makes the TAST a less valid test since scores may represent only fine-tuned test-taking ability, not real, spontaneous language ability. However, the training opportunities also make it popular with examinees since they can feel more in control of their test results.

Table 3.3

Summary of TAST Test Content

Prep/Speaking time	**TAST task**	**Addresses this TA task**
15s/45s	Descriptive	Teaching content
30s/60s	Summary	Teaching content
20s/60s	Argumentative	Justifying grades
20s/60s	Following instructions	(doesn't address a TA task)

Both a strength and a weakness of the TOEFL iBT is that its construct was developed to test integrated skills. So, as described above, in the TAST, examinees speak in response to spoken and/or written input. This is a strength because it acknowledges the inter-dependence of language skills, and the complexity of even basic language tasks. However, it also makes it difficult to know exactly what skill is being tested. Perhaps the tester struggled to understand and remember all the details in the spoken prompt and so gave an inappropriate spoken reply. Or, perhaps the tester is a slow reader, and couldn't read the entire written prompt and so was unprepared to speak.

The TAST is scored by two raters (with a 3rd rater brought in if the scores do not agree) by evaluating each task on a scale from 0-4 (See Appendix B). The average of all 6 tasks is then converted to a scaled score of 0-30. Scores have been extensively tested for inter-rater reliability, as well as to establish the validity of analytic rubric versus a more holistic rating method (Sawaki, 2007). The overall score mean of participants in this study on the TAST is 20, but in fact the distribution of TAST scores in this study is bi-modal, with means of 19 and 23, with 2 outliers, one less than 9 and one less than 30. The standard deviation reveals that there is much greater variation on the TAST scores (sd=2.79) than on the TBEST (sd=.83) (Table 3.4).

Table 3.4

TBEST and TAST Descriptive Statistics

	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
TBEST	335	3.57	9.94	7.01	.83
TAST	193	9.00	30.00	20.84	2.79
Valid N (listwise)	193				

3.6.3 Description of the TCE. Teacher Course Evaluations (TCE) measure ITA teaching effectiveness. The TCE is administered at the end of every UA class, every semester. There are twenty versions of the TCE available: a short and long form, and then various discipline specific versions. However, there are 11 core questions that are on all forms, and four of those specifically address overall teaching effectiveness ("TCE Guide," p. 6). This present analysis considers those four questions for ITAs:

1. What is your overall rating of this instructor's teaching effectiveness?
2. What is your overall rating of this course?
3. How much do you feel you have learned in this course?
4. What is your rating of this instructor compared with other instructors you have had?

The Office of Institutional Research and Evaluation (OIRE) reports that the first three questions on instructor, course, and amount learned are usually highly inter-correlated ("TCE Guide," p 6). The fourth question of comparison is generally "problematic" since students tend to interpret the comparison element differently: comparison to all other classes, only to classes in the major, classes at the UA or also at other institutions. This factor was considered, but not supported by the TCE results of ITAs in this study (see section *4.3.2.3 Correlations* to see the TCE results). However, answers to these four questions are individually correlated, so that individual question variation does not effect the overall correlation.

Systematic variation in the ratings has been noted in some factors, and these are considered during my analysis.

1. Disciplinary Differences: Humanities and Fine Arts courses tend to be rated higher than courses in Physical and Applied Sciences. Therefore ITAs' TCE reports were not compared across discipline.
2. Course Status: Electives and courses in majors tend to be rated slightly higher than general education requirements. ITAs typically teach general education courses, so if that bias exists, it should be consistent across ITA scores.
3. Years of Teaching Experience: Instructors with less than one year of experience tend to receive the poorest scores. Most ITAs in this study have not taught previously, or at least have not taught in the United States. ("TCE Guide," p. 19-20)

3.6.4 Description of the ITA survey. ITAs who passed the TBEST assessment, were later surveyed using an online, fixed-response survey after teaching for one or two semesters. 119 ITAs participated in the survey. One purpose of the survey was to identify the speaking tasks that ITAs perform in their teaching positions, in order to triangulate with the speaking tasks that were identified by the TBEST re-development team, and the tasks that are evaluated on the TBEST and the TAST. The other purposes relate to the pedagogical implications of this dissertation: ITA training and ITA teaching evaluation.

The survey includes questions about teaching, training, and evaluation. Each question asks participants to rank the possible answers on a Likert scale. Some demographic information is also included about their teaching background and class sizes. The final question asks about elements of their overall satisfaction as ITAs *(See Appendix C: ITA follow-up survey questions).*

CHAPTER 4

ANALYSIS OF RESULTS

The overarching goal of this dissertation is to compare alternate assessment tools in order to assist decision makers in choosing the most valid manner of screening potential ITAs. The application of the chosen assessment tool is ITA placement along a hiring scale that denotes recommended levels of student contact, 1) Full instructional responsibility, 2) Full instructional responsibility in the ITAs' native language (as foreign language instructors), 3) Limited instructional responsibility as graders or lab leaders, or 4) Not eligible to be an ITA. This goal is distinct from a general or generic knowledge of English as it involves primarily spoken communicative competence with a specified audience, undergraduate students.

As I evaluate the TBEST and TAST, I search for a preponderance of evidence for assessment validity that indicate the most appropriate tool for evaluating potential ITAs. In this chapter, the specific evidences of assessment validity that I examine are:

1. Evidence of Domain (Content) Validity: Which test, the TBEST or the TAST most closely measures the actual skills needed to be an ITA?
2. Evidence of Predictive Criterion Validity: Which test, the TBEST or the TAST, is more valid in predicting ITA teaching success based on end of semester student evaluation (TCEs)?

Following the analyses of these points of evidence, I review the results of a follow-up survey of ITA impressions about the ITA training and evaluation process. Reviewing the results of this survey allow me to place the language assessment and hiring process within its larger context and direct attention toward suggestions for improvement of the overall ITA system.

4.1 Parallel Test Forms

Before analyzing evidence for validity for the TBEST and TAST, I confirmed that they are not parallel forms of assessment. To do this I correlated TBEST and TAST scores for those ITA examinees who have taken both exams. The Pearson's correlation (r=.592) is a medium correlation, which implies similarity of assessment, but not so high as to suggest that the TBEST and TAST are parallel forms of assessment (Figure 4.1).

		TBEST	TAST
TBEST	Pearson Correlation	1	.592**
	Sig. (2-tailed)		.000
	N	335	193
TAST	Pearson Correlation	.592**	1
	Sig. (2-tailed)	.000	
	N	193	193

Figure 4.1. Pearson Correlation of the TBEST and TAST
** Correlation is significant at the 0.01 level (2-tailed).

4.2 Evidence of Domain Validity

Studying evidence of content validity is the primary endeavor of this study. According to Reynolds (1997), domain- or content-based validity evidence is preferred for establishing the validity of academic achievement and on tests used in the selection and classification of employees, making this a justified route of inquiry for examining an ITA hiring prerequisite. Furthermore, following Anastasi and Urbina's (1997) instruction, I ask: "Does test content provide a representative sample of the domain being measured?" (p. 116). In order to identify skills in the ITA domain, I first look internally at an explanation of TA responsibilities at the UA as determined by a panel of experts and stakeholders in the ITA hiring process. Then I check the accuracy of the committee by conducting a follow-up survey of new ITAs, asking about their most common speaking tasks as ITAs.

As I noted in Chapter 2, the tasks identified by the expert panel are: teaching general content, explaining discipline specific theories, justifying grades, and classroom/lab management. These are the tasks that were agreed upon across disciplinary lines, recognizing that other required skills may be specific to individual disciplines.

Table 4.1

"What types of speaking tasks does your TA position require?"

Task types*	**Average response**
Explaining /Elaborating	4.45
Describing something	4.26
Giving suggestions/advice	3.95
Justifying grades/homework points	3.81
Re-stating information in a simple way	3.75
Giving step-by-step instructions	3.68
Giving formal presentations	2.99
Disciplining students	2.63
Reading aloud from a book	1.90

N= 110, Rated on 5pt Likert scale, 5=most frequent, 1= least frequent
* To avoid conflating terms, see Underhill (1987) for definition of the relevant task types, and the analysis in the following section.

According to the results of the follow-up survey (Table 4.1), the most common speaking tasks of ITAs are: explaining, describing, giving suggestions/advice, justifying grades/homework points, and re-stating information in a simple way. The survey question was deliberately worded differently than the tasks on the TBEST to avoid data contamination. Between these two groups, there is agreement that the most essential speaking tasks include Explaining, Describing, Justifying grades, and Re-stating complex information in simpler terms. Giving suggestions or advice is the only task that ITAs identified that is not within the domain of either the TBEST or the TAST.

4.2.1 Comparative domain analysis between TBEST and TAST. In this section, I compare each of the four primary ITA speaking tasks that were identified by the committee and by ITAs to the tasks on the TBEST and TAST: Describing, Explaining, Paraphrasing, Justifying, and based on the ultimate application of this assessment, Classroom/Lab management. This comparison reveals the similarities and differences between the assessment tools, an overview is in Table 4.2.

Table 4.2

Comparison of TBEST and TAST Content to the ITA Teaching Domain

ITA Duties	TBEST Content	TAST Content
Describing, Teaching general content	Descriptive Task	Descriptive Task Summary Task
Explaining	Comprehensive (Video) Task	*none*
Paraphrasing	Interpretive Task	*none*
Justifying grades	Argumentative Task	Argumentative Task
Classroom/Lab Management	Comprehensive (Video) Task	*none*
		Following Instructions Task *Does not correlate to an ITA duty*

4.2.1.1 Descriptive task. According to Underhill, the benefit of using a descriptive task in spoken language assessment is that in choosing a topic that is familiar to everyone, the examinees are able to “produce connected discourse on a given topic, but allows considerable freedom of choice of expression without requiring extensive preparation" (Underhill, 1987, p. 69). In this particular assessment context, on a daily basis ITAs must be able to describe a wide variety of information in class. It is not surprising that both the TBEST and the TAST contain tasks that measure general descriptive ability. This is a foundational and high frequency communicative task and therefore descriptive tasks are present on most oral assessment tools. On both the TBEST and the TAST, examinees are asked to describe specific experiences, people, or favorite items. This general descriptive ability is used to approximate ITAs ability to describe foundational elements of their courses, assignments, and answers to student queries. The TAST also contains a summative task, in which examinees summarize either a reading passage or a lecture excerpt. There may be some overlap in the speaking skills of description and summary.

4.2.1.2 Explanation task. Underhill (1987) defines the content of an explanatory task saying "for learners who need English for academic or professional purposes, the topics for explanation are chosen from a small number of specialist subjects" that require description or explanation using professional terminology, but not to be "so difficult as to become a test of technical knowledge rather than proficiency" (p. 71). In the case of ITAs, it is important that they are familiar with classroom management strategies, basic pedagogy, and common student disciplinary issues. These issues are presented in the TBEST comprehensive video task of a problematic classroom scenario. Examinees are instructed that this is a comprehensive task, that is, that they should describe the situation that they observe, interpret the actions of the teacher or students, and explain their reaction to the scene. There is not a comprehensive explanatory task of an academic or pedagogical topic on the TAST.

4.2.1.3 Paraphrase task. Explaining theories is a slightly different task than basic content teaching or giving a general explanation. Discipline-specific theories often employ metaphor or abbreviated language to represent a more elaborate concept. For example, in SLA, common theoretical terms include the *ZPD* (Vygotsky's Zone of Proximal Development), which in turn feeds into the pedagogical concept of *scaffolding*. Explaining theories like these requires an ITA to explain and disambiguate the symbols or references in a theory and make them comprehensible to a novice learner.

In the TBEST, this speaking ability is assessed by asking examinees to explain the possible meaning of proverbs. Proverbs were chosen which required the least cultural context to make this task fair for all participants. The TAST does not have test content that evaluates this speaking skill, though there is a somewhat similar task of simple summary in which the examinees summarize a long written passage in their own words, although summary does not require the skill of disambiguation.

4.2.1.4 Justification task. The aim of this task is to allow examinees to demonstrate their ability to "use the language effectively to justify a position and not just state it" (Underhill, 1987, p. 70). All teachers have to be able to justify grades, and this may be even more true with ITAs. Especially if students doubt the authority of the ITAs, students may feel emboldened in challenging ITA grading decisions. Therefore, ITAs must be able to make an argument and support their position with convincing reasons in English. Both the TBEST and TAST contain an argumentative task that can be used to measure this speaking ability.

4.2.1.5 Instruction task. A task that is on the TAST, but not on the TBEST, calls for examinees to listen to or read a set of instructions, and then perform the task that is described. For example, they may be instructed on the proper components to leave in a voice mail message, and then be given a scenario prompting them to leave a particular type of message. This task falls into a general communication category that is appropriate for general language testing, but does not necessarily apply to the needs of ITA hiring assessment based on the results of my ITA task survey.

4.2.1.6 Classroom management. The most difficult teaching construct to assess is classroom management. The difficulty in assessing classroom management is that the assumed ideal assessment would include a performative element, however, once performance is actually introduced, the reliability of the assessment decreases. In addition, some research reports that this is the least efficient way to measure classroom management ability due to all the participants that are needed to create a mock class (Plakans & Abraham, 1990, p. 78). Additionally, Plakans and Abraham found that some of the evaluators who were serving as mock-students reported that they interacted significantly different over the course of the evaluation day, being much more challenging with some ITAs than with others (p. 78).

To prevent these possible irregularities, but to still measure ITAs' awareness of classroom management issues and strategies, the TBEST uses videos of various teaching scenarios and asks examinees to watch and then comment on the scene. Not only does this assess the examinees' ability to create explanatory discourse as described above, but it also demonstrates their awareness of classroom management issues. There is no measurement of classroom management assessed in the TAST.

In this section, I have analyzed the test content in both the TBEST and TAST specifically in light of the speaking tasks that ITAs must perform in the classroom. I found that, by design, the TBEST assesses the primary skills in the ITA teaching domain, while the TAST assesses general language and only two of the five identified ITA speaking tasks. A visual summary of the test content on the TBEST and TAST is presented in Table 4.2. In the next section, I transition from domain analysis to statistical analysis of predictive criterion evidence.

4.3 Evidence of Predictive Criterion

The second goal of this study is to evaluate the predictive criterion for evidence of assessment validity for both the TBEST and the TAST. Concurrent criteria are assessments that are administered at the same time and Predictive criterion compares preliminary scores with a measurement taken after the passage of time (Reynolds & Wilson, 2006, p. 127). In this phase of my research, I look at predictive criterion, to see which hiring tool, the TBEST or the TAST has a higher correlation to successful teaching based on end-of-semester student evaluations.

4.3.1 Limitations. There are limitations to this analysis that I must acknowledge before proceeding. In order to conduct this predictive correlation analysis, the population that I examined had to have taken both the TBEST and the TAST, and to have taught as the primary instructor (or in one case as a team-teacher) and to have been evaluated using the TCE. In the roster of 335 TBEST examinees there were 195 who had taken both exams. Of those 195, 122 earned a score that qualified them to be a primary instructor, but only 11 had TCE scores.

There are multiple possible explanations for this lack of TCE scores. The most likely explanation is the type of position given to ITAs. In my follow-up survey, only 20% of respondents reported that they were the primary instructor of a course, the majority (61%) were graders or lab/discussion leaders and others were hired as research assistants or as private tutors. Therefore, the majority of ITAs are hired into positions in which formal performance/student evaluation is not required or conducted. Other possible explanations are that departments may instead rely on some form of internal evaluation, or scores may have accidentally not been reported.

With such a small population, care should be used when interpreting the data, and these statistical results are not appropriate for making an assessment decision in this particular case. However, it is still important to document emerging trends and demonstrate the steps of evaluating evidence of criteria correlation for future studies of this nature.

4.3.2 Criterion data. TCE (Teacher Course Evaluations) correlations are the primary evidence that I consider for criterion validity, as they are the university standard for evaluating all UA instructors (ITAs, GTAs, Professors). The TCEs come in several different versions, depending on the course, but the questions are generally a mix of questions about the teacher's effectiveness, the course materials, and overall student demographics.

4.3.2.1 Participants. Of the 11 TCE scores that are reported, eight are from first and second year foreign language courses, two are from a civil engineering course and lab and one from a communication (speech) course. The ITAs reported here taught between 1-9 courses during the three semesters of this study. While I previously wrote that it would be necessary to analyze reports separately by discipline, there was not a significant difference in the TCE scores between the humanities and the science courses. In addition, contrary to the presupposition that students rate ITAs lower than proficient English speaking TAs, there is no significant difference between TCE scores of this ITA population and the average scores of the comparison group (Table 4.3). In fact, ITAs in my survey scored slightly higher than the comparison group on Questions 2 and 4.

Table 4. 3

Descriptive Statistics of TCE Score Results

	N	**Min**	**Max**	**Mean**	**Std Dev**	**Comp Mean***	**Comp SD**
Q1**	11	3.18	4.50	3.97	.38	4.09	.49
Q2	11	2.90	4.43	3.75	.49	3.70	.47
Q3	11	2.72	4.13	3.45	.46	3.82	.50
Q4	11	3.07	4.43	3.73	.49	3.72	.54

*Comparison group TCE scores from 3-credit, lower-division, undergraduate courses ("TCE Comparison," p. 1).
**All questions rated on 5-point Likert scale. 5=positive response, 1=negative response
Question 1: "What is your overall rating of this instructor's teaching effectiveness?"
Question 2: "What is your overall rating of this course?"
Question 3: "How much do you feel you have learned in this course?"
Question 4: "What is your rating of this instructor compared with other instructors you have had?"

4.3.2.2 TCE questions. The responses to four questions on the short form TCE are used because they are on all variations of the TCE and they most closely focus on the evaluation of ITAs' teaching ability, as opposed to the components of the specific course. These four questions are the "overall" questions that are recommended by OIRE (The Office of Institutional Research and Evaluation) for "performance appraisal because they are applicable across the wide variety of teaching styles and course formats" ("TCE Guide," p. 6). The four questions that I analyze are:

1) What is your overall rating of this instructor's teaching effectiveness?

2) What is your overall rating of this course?

3) How much do you feel you have learned in this course?

4) What is your rating of this instructor compared with other instructors you have had?

4.3.2.3 Correlations. A correlation matrix between the four TCE questions and the TBEST and TAST results allows me to view all potential correlations (Table 4.4). The correlations between the TCEs and the assessment tools are small to medium, with large correlations only occurring inter-item, between the TCE questions. Questions 1 and 3 (concerning instructor effectiveness and amount students learned) and Questions 2 and 4 (concerning the rating of this course and comparing this instructor to others) share the largest correlations, r= .896, .739, respectively. The overall low correlations throughout the rest of the matrix could merely reveal that the questions on the TCE do not directly ask for students' opinion of ITAs' speaking ability, and that is what is explicitly tested on the TBEST and TAST.

Table 4.4

Correlations of TCE Results with TBEST and TAST Scores

		Q1	Q2	Q3	Q4	TBEST	TAST
Q1	Pearson Correlation	1	.547	.896**	.575	.237	-.376
	Sig. (2-tailed)		.082	.000	.064	.482	.254
	N	11	11	11	11	11	11
Q2	Pearson Correlation	.547	1	.550	.739**	-.350	-.539
	Sig. (2-tailed)	.082		.080	.009	.292	.087
	N	11	11	11	11	11	11
Q3	Pearson Correlation	.896**	.550	1	.430	.203	-.311
	Sig. (2-tailed)	.000	.080		.187	.549	.351
	N	11	11	11	11	11	11
Q4	Pearson Correlation	.575	.739**	.430	1	-.147	-.294
	Sig. (2-tailed)	.064	.009	.187		.666	.381
	N	11	11	11	11	11	11
TBEST	Pearson Correlation	.237	-.350	.203	-.147	1	.416
	Sig. (2-tailed)	.482	.292	.549	.666		.203
	N	11	11	11	11	11	11
TAST	Pearson Correlation	-.376	-.539	-.311	-.294	.416	1
	Sig. (2-tailed)	.254	.087	.351	.381	.203	
	N	11	11	11	11	11	11

**. Correlation is significant at the 0.01 level (2-tailed).

Note: In this population (n=11) TBEST scores have a SD of .80 and TAST scores have a SD of 2.10, see Table 3.4 for complete descriptive statistics, and Table 4.5 for descriptive statistics of just this population.

Table 4.5

Descriptive Statistics of TBEST and TAST data for those with TCE scores

	N	Minimum	Maximum	Mean	Std. Deviation
TAST	11	18	23	20.27	2.10
TBEST	11	5.16	7.72	6.75	.80

Next, I examine the correlations between TCE scores and the TBEST and TAST. First, the TBEST has a small, insignificant, correlation to TCE scores, on Questions 1 (r=

.237) and 3 (r=.203), the questions on overall teacher effectiveness and the amount the students felt that they learned. On Questions 2 (r=-.350) and 4 (r=-.147) there is a negative (inverse) correlation. Secondly, the TAST is negatively correlated to TCE results, an increase in X correlated to a decrease in Y, on all four TCE questions, r= -.376, -.539, -.311, -.294, respectively. While these results are to be interpreted cautiously due to the small sample size, these results could suggest that the TBEST is more valid in predicting ITAs' teaching effectiveness. A follow-up study should be conducted with a larger pool of TCE results to test these results.

In this section, I analyze the evidence for predictive criterion validity between the TBEST and TAST and the TCE reports. Due to a small sample size (n=11) these results must be interpreted cautiously and the correlations should not be used to make decision unless a follow-up study is conducted with a larger population. The preliminary results here show that the TBEST tends to have a small correlation to TCE questions on teacher effectiveness (r= .237) and the amount students feel they learned (r=.203), and that the TAST tends to be negatively correlated to all four TCE scores. These results will be further discussed in Chapter 5. Regardless of the correlations, as Reynolds (1997) wrote, domain-validity is key for employee selection assessment and therefore it is more valid to consider the domain analysis for making assessment decisions for this ITA assessment context. Having discussed the Domain and Criterion evidences for validity, I now turn my attention to the broader context of the ITA situation, focusing on ITA preferences in training and course evaluating.

4.4 Qualitative Contextualization

Having analyzed the domain validity and predictive criterion in the previous sections, I turn my attention to the contextual information that ITAs provide in their follow-up survey responses. In this section, I examine the survey responses on a range of questions about training and evaluation procedures and discuss narrative replies about ITA issues that are important to the respondents. While discussing training options, I look at the issue from both the perspective of what ITAs would like included in a training course, and also the teaching variables that they found difficult as ITAs. In the section on evaluation, I discuss ITA responses about how they would like to be evaluated, and what they think students prioritize in evaluating ITAs. I also consider the need for demonstrating useful application of evaluation to ITAs for their own professional development. Finally, I summarize issues of concern that ITAs express about general hiring procedures.

4.4.1 Participants. All 335 ITAs who participated in the TBEST during the scope of this survey were invited to take a follow-up survey to measure their impression of the overall ITA hiring, training, and evaluating process at the UA. The response rate was 36% (n=119). Of those responding, 109 passed the TBEST and are eligible to be an ITA. Of the 10 who did not pass, 7 plan to re-take the TBEST in the future and the remaining 3 decided not to become ITAs. In describing their current TA assignment, 41% serve in a medium-sized class (20-40 students), and the remainder are nearly evenly split between small (less than 20 students), large (40-100), and extremely large (more than 100 students) classes. The majority (37%) of these ITAs are lab or discussion leaders, 24%

are graders, and just 20% are primary instructors. The remaining ITAs serve as team teachers, research assistants, or individual tutors.

4.4.2 ITA training. An important lesson from the TCE analysis is that it takes more than speaking ability to be an effective teacher. In this section, I examine ITA attitudes about TA training possibilities. In the follow-up survey, 40% of ITAs are interested in participating in a TA training course, 60% feel that the current 2-day TA orientation (GATO) and online policy modules (TATO) are sufficient. This may be more reflective of the overall workload that ITAs struggle with than the actual perceived need for training.

Table 4.6

Summary of ITA Class Sizes and Instructional Roles

Class size	Sm: <20	Med: 20-40	Lg: 40-100	Ex Lg: 100+	N=
	18	41	22	20	101

ITA's Role	Research Assistant	Primary Instructor	Lab/Discussion Leader	Team Teacher	Grader	Other	N=
	4	20	37	6	24	11	102

Opinions are split on when the training should ideally take place. The majority (38.6%) want a course prior to being a TA. While 24.6% prefer a training course before they had to take the speaking test and 26.3% are interested in a course that runs concurrent to their first semester of teaching. Just 10.5% said they prefer to take the training course after their first semester of teaching. Following the majority's will would improve administration of the testing and hiring process, as ITAs would have their language qualifications cleared a semester in advance of actually teaching. Requiring ITAs to take a semester long training course prior to teaching would also allow them to become even more proficient through target language practice in the training course. As outlined in Chapter 2, this option has precedent at other universities with established ITA training programs.

The top 4 elements that ITAs want included in the training are: speaking fluency practice, general teaching strategies, presentation skills, and pronunciation practice (Table 4.7). The bottom 4 training topics are: meeting with a faculty mentor, developing syllabi/assignments, having a class observed, and cultural teaching tips. From these results it seems that ITAs have more concerns about their language ability than about their actual teaching ability, as the tasks with the higher ratings tend to be primarily speaking tasks, while the lower rated items are primarily teaching tasks.

Table 4.7

"What would you like an ITA training course to include?"

Training elements	Average Rating	Speaking/ Teaching Task
Speaking fluency practice	4.19	S
General teaching strategies	4.05	T
Presentation skills	4.04	S
Pronunciation practice	4.04	S
Practice teaching scenarios with your class material	3.84	T
Cultural teaching tips	3.66	S/T
Observing other classes	3.64	T
Faculty mentor meetings	3.41	T
Developing syllabus/assignment sheets	3.39	T
Having your class observed	3.38	T
Powerpoint demonstrations	3.23	T

N=61 (not all respondents answered this question)
Rated on a 5-point scale, 5=most useful, 1= least useful

Looking at this issue from another perspective, I ask what elements of teaching were most difficult for the ITAs (Table 4.8). While the respondents were very reluctant to admit difficulty (the most common answer had an average score of 2.92 out of 5), the tasks with the highest average ratings were teaching skills, not speaking skills: leading student discussion, knowing when students understood/were confused, and planning lessons, and grading. These results conflict with the skills that ITAs identified as training priorities (in Table 4.7), and I believe this reveals that ITAs, like students, place too much focus on improving speaking ability as a panacea to improving their teaching ability.

Table 4.8

"What parts of teaching at the UA were difficult for you?"

Tasks	Average Rating	Speaking/ Teaching Task
Leading student discussions	2.92	S/T
Presenting lessons in class	2.63	S
Knowing when students understood/were confused	2.60	T
Planning lessons	2.52	T
Grading	2.49	T
other: (no tech in class, time management)	2.40	
Creating assignments	2.37	T
Creating tests	2.35	T
Explaining assignments to students	2.27	S
Solving conflicts between students	2.27	T
Answering student questions in office hours	2.26	S
Explaining grades to students	2.21	S
Dealing with cultural differences with students	2.21	S
Using D2L (online classroom management tool)	1.89	T
Using classroom technology	1.84	T

N=106. Rated on a 5-point scale, 5=most difficult, 1=least difficult

As Gorsuch (2006) points out the two directions that ITA training can take are teacher training, which is best implemented by individual departments and alongside proficient English speaking TAs (p. 91). The other option is language education, which can be delivered through a centralized program for multiple academic departments, taught by ESOL or SLA specialists (p. 91). Gorsuch is quick to point out that these options are not an "either/or proposition" (p.91) just that each university chooses to integrate these approaches differently. Whatever the plan, Gareis (2004) acknowledges the importance of training saying that "nonnative instructors often enter a plateau of second language proficiency and teaching competence after a few years of residence, and are able to improve upon this level only with conscious effort and some assistance" (p. 46). This means that by the time ITAs graduate and seek university employment, their teaching style is likely engrained. To make long-term improvements on undergraduate education, ITAs need training before they become international faculty.

4.4.3 ITA evaluation. Performance reviews are an integral part of any academic career, so they should also be well integrated into the ITA experience. Based on the survey, ITAs tended to favor more traditional methods of evaluation rather than alternative assessment options. By wide margins, the ITAs in this survey prefer end-of-term student evaluations (4.07 on a 5-point scale), then mid-term student evaluations, and seeing these results at that time (3.97), and then advisor observation (3.93). The less popular methods are peer evaluation (3.37), student passing rates (3.37), teaching material evaluation (3.36), and self-evaluation (3.21).

These results reveal ITAs' attitudes toward the current evaluation system. The most noticeable trend is the preference for student evaluation. I believe this reveals that there is a preference for evaluation from students who are in an inferior power position because ITAs can dispute the students' evaluation if it is unfavorable. Next, by preferring advisor evaluation to peer evaluation, it shows that they value expert input over (presumable) novice input. Eventually, if ITAs go on to pursue academic careers, they will need a recommendation from their advisor, so it is advantageous to get their input early and often.

It is not clear from these results what impact ITAs see on their teaching from the TCE evaluations, though since such a minority (approximately 10% of this sample) actually receive TCE reports, it is safe to assume that most ITAs do not see an immediate or actionable result that comes from regular evaluation. This is a disservice to ITAs since self-reflection and goal-setting as professional development are often components of the hiring or tenure promotion process in disciplines that emphasize teaching, and ITAs should become accustomed to this process.

A second evaluation question has implications for training (Table 4.9). When I ask what ITAs think students value most when evaluating ITA teaching ability, their response shows a picture of a well-rounded instructor. The top criteria include a mix of speaking, teaching, and personality (culturally influenced) traits, not primarily speaking or teaching as in other iterations of this question. These results continue to make the case for a comprehensive training course.

Evaluation can be improved by emphasizing the formative value that it can have on their teaching, rather than the current system that is commonly seen for its assumed summative implications, of qualifying ITAs to keep their jobs. Furthermore, ITA attitudes about what students value can be utilized to have positive washback on the ITA training process.

Table 4.9

"When students evaluate ITAs, what do you think students value most?"

Evaluation Criteria	Average response	Speaking, Teaching, or Personality Traits
Fairness in Grading	4.42	T
Feeling comfortable asking ITA questions	4.40	S
ITA's Content knowledge	4.34	T
Feeling respected by the ITA	4.23	P
Clear pronunciation	3.79	S
Easy Grading	3.50	T
Humor	3.35	P

N= 103. Rated on a 5-point scale, 5= most valued, 1= least valued.

4.4.4 Overall impressions. Beyond getting feedback about what kind of teacher training would be desirable and how evaluations could be more useful, the survey also allowed me to measure the ITAs' satisfaction with their own teaching. The results of the survey depict a largely positive ITA experience and challenge many stereotypes about the ITA experience.

The majority of respondents feel that being an ITA is good training for their future and economically beneficial. In the classroom, ITAs report that their students respect them and treat them the same as an American TA. ITAs in this survey do not seem to think that being a TA has a negative impact on their own progress toward graduation or on their own preparation for the classes they are taking, though it is time consuming. One ITA comments that, "It takes more time for an ITA to do the same task for an American TA."[8] Another stated, "Being a TA is a great chance to get experience at the American College level, however, it is extremely time-consuming and sometimes that is not really appreciated by our professors in grad school or at the university in general."

[8] All comments from the ITA survey are directly quoted, including any errors.

These points of view are widespread as Figure 4.2 show, and should be considered when developing ITA training. The training needs to be practical and hands-on, and not overly time-consuming. In addition, since so many ITAs see their time as TAs as *training* for their future occupation, departments have an obligation to provide that training, or make their intentions not to emphasize training, and not just view ITA positions and funding as a primarily financial way to attract international scholars, though the funding is clearly important as well.

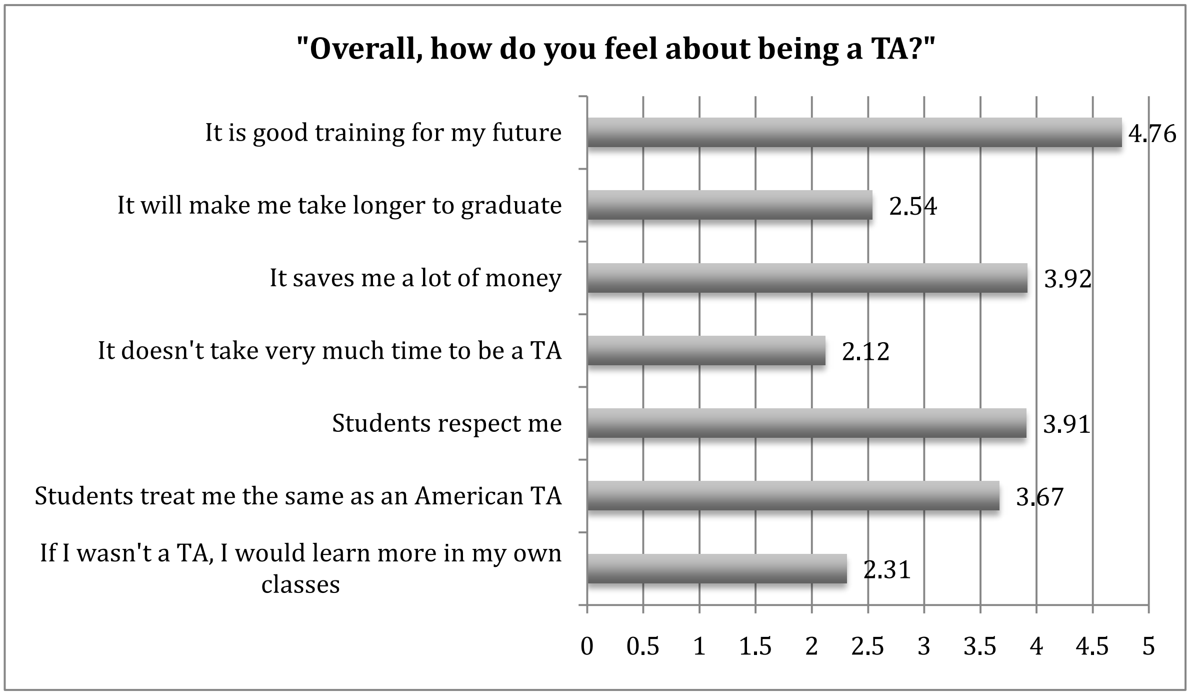


Figure 4.2. Overall ITA Satisfaction
N= 103. Rated on 5-point scale, 5=strongly agree, 1=strongly disagree.

There was discontentedness among some ITAs for having to take this speaking test before becoming an ITA. One ITA commented: “I think student from certain foreign country that has English as their FIRST language doesn’t need to take TBEST. It is completely a waste of time and money for me." It is likely that this graduate student is from India, since that is one of the largest countries represented in this survey population, and English is the first language of many Indian students. This comment brings up an important discussion of World Englishes, and which English varieties ought to be privileged in American academia.

Another student reported discontent with the hiring process at the UA saying that he/she taught a lab successfully for 2 weeks before taking the TBEST and scoring a 6. At that point the department took away this student's position and funding for the semester leaving the graduate student very stressed and depressed until retaking the TBEST and passing with a 7 the following semester.

My frustration with this scenario is that numerous points of protocol were broken on the part of the department, to the detriment of the ITA. First, no department should hire an ITA before the hiring requirements have been met, so this student should have had other financial arrangements for the first semester. Secondly, the TBEST score of 6 is sufficient for lab leadership. A 7 is only required for ITAs who are the primary instructors, which I learned is a minority of ITAs.

While this is the only response that hints at this problem, other ITAs did mention similar situations when coming back to re-take the TBEST, and I highlight it here to point out the importance of communicating the intended application guidelines with all stakeholders in any assessment process. ITAs are a particularly vulnerable population that should be protected since they do not have the cultural or social resources to support themselves in cases where they may be disadvantaged or even taken advantage of. Creating a screening process for ITAs is important, but that process must be properly and consistently applied.

In support of the testing process, one ITA reports, “I would emphasize the importance of English speaking test. Some of my other students told me stories that some other TA’s were not able to help them with material due to language barrier.” However, another ITA puts the responsibility on the students, saying that “[f]or good students, ITA’s accent is never a problem” and sometimes students do have difficulty understanding, but that is because of “their own lack of concentration, however it is just easy for them to blame ITA’s accent.” It is important to ITAs that overall ITA quality standards are upheld, because the negative impact that the whole population feels when students have experiences with un(der)qualified ITAs, and negative stereotypes are spread about ITAs.

Overall ITAs in this survey appear to conflate speaking fluency with successful teaching. One respondent says, “I think for me, English is still the most important. If I can speak English fluently, I'll be a good TA.” And another says, “I think the most important think is that an International TA is able to communicate with the students in English.” Another laments, “sometimes the students do not get your pronunciation. The bad thing is that they never tell you that until the year end evaluation even though you ask them about it in the class.” I think that overall professionalism, content knowledge, and pedagogical strategies are overlooked in ITA assessment and evaluation, and therefore it seems less important in ITA training, and in ITA self-reflection.

In this section, I examined the holistic context of ITA training, evaluating, and

ITAs general feedback on their experiences. The survey shows that while adding training isn't favored by the majority (but rather, 40%), ITAs were able to identify weaknesses in their teaching ability, and therefore, a comprehensive training course that emphasizes both pedagogy and speaking should be developed. The responses to evaluation questions suggest that ITAs do not currently see the usefulness or impact of evaluation in their classes. Furthermore, current evaluation seems to be lacking, since methods like TCEs do not evaluate the skills that ITAs value. Finally, language testing is seen as important for quality control, but the issue of World Englishes needs to be addressed. Many of the suggestions hint toward possible next steps in research and in training development, which are discussed in the final chapter.

4.5 Conclusion

In this chapter I pursued two research questions in search of evidence for Domain and Criterion validity:

1. Evidence of Domain (Content) Validity: Which test, the TBEST or the TAST most closely measures the actual skills needed to be an ITA?
2. Evidence of Predictive Criterion Validity: Which test, the TBEST or the TAST, is more valid in predicting ITA teaching success based on end of semester student evaluation (TCEs)?

Regarding the first research question, the TBEST has greater evidence of domain validity in that its test content addresses all five critical skills (four speaking skills and classroom management) that were identified by the stakeholders committee and by ITAs. The TAST, on the other hand, only assesses two of the critical skills: description and making an argument.

Regarding the second research question, the answer is less clear-cut. The TBEST has somewhat higher evidence of predictive criterion validity though neither the TBEST nor TAST has a significant level of correlation to TCE scores. The correlation between TBEST and TCE is small on questions of teacher effectiveness and the amount students learned (r=.237 and r=.203). These results are complicated in that, on the other two TCE questions on course quality and the rating of the instructor compared to others, the TBEST has negative correlations (r=-.350 and r=-.147). However, the TAST has negative correlations to all four TCE questions, r= -.376, -.539, -.311, -.294, respectively. While these results are to be interpreted cautiously due to the small sample size (n=11), these results could suggest that the TBEST is more valid in predicting ITAs' teaching effectiveness, but only if they are confirmed with a study of a larger population.

The third goal of this analysis was to place the process of choosing the appropriate assessment tool within a larger context that ITAs experience while applying to teach as an ITA. I found that most TAs have positive reactions to the TBEST, even with the time and monetary costs, as they see the implications of not having adequate English speaking ability as an ITA. Anecdotally, I have been told that they understand the frustration of undergraduates (especially international undergraduates) who cannot understand their professors or TAs. I think that because they are themselves international students who have come to the U.S. to attend graduate school, they understand the frustration of not getting the educational opportunities that they expected when they came to an American university.

With that said, there are deficiencies in the ITA system overall. There must be more effort to standardize the hiring and testing process, so that ITAs know their funding, ideally leaving their home country. There does need to be more quality training, but with the acknowledgement of the time constraints that ITAs are already working within. Evaluation of ITAs needs reimagining, so that all ITAs have the benefit of receiving regular feedback. In the final chapter, I discuss in greater detail future directions that I see arising from these results.

CHAPTER 5

DISCUSSION

This chapter contains the summary and potential application of my research. I then reflect on the ITA hiring process and future TBEST research. I first briefly summarize the results of this dissertation research to provide context for my discussion of possible applications of my findings. Specifically, I consider the implications of this research for ITAs' hiring, training, and evaluating procedures and for future studies of this nature. Finally, I reflect on how my own perceptions were challenged over the course of this research.

5.1 Summary of Results

In this section, I summarize my findings regarding the two research questions at the foundation of this study. Then I outline the predominant findings from the ITA follow-up survey.

5.1.1 Domain analysis. The most important task in creating a domain-referenced assessment tool is identifying the appropriate test content that provides an authentic, representative sample of the domain being measured (Anastasi & Urbina, 1997, Reynolds, 1997).

Deliberative steps were taken to ascertain the key speaking tasks that underlie the domain of undergraduate instruction. Before repurposing the TBEST for ITA evaluation, the ITA instructional domain was defined through a consulting process with department heads, ITA advisors, and graduate TAs, deciding on the most common speaking tasks in the TA setting. Secondly, the assumptions of the panel were confirmed by surveying new ITAs (n=119) about their most common speaking tasks in the classroom.

The results showed agreement between these two groups. Five key tasks emerged: Teaching content, which can be broken down into 1) Descriptive and 2) Explanatory skills, 3) Paraphrasing complex discipline-specific theories which often contain metaphors and abbreviated language, 4) Argumentation/Justification of grades, homework answers, points on quizzes, and 5) Knowledge of Classroom/Lab Management strategies. The only skill that was a priority of ITAs, but not included in this final task list was giving advice or suggestions. This could be another task to consider including in future ITA assessment tools.

Based on these five speaking tasks, I analyze the content in both the TBEST and TAST and find that the TBEST, by design, includes all 5 variables in its assessment content. The TAST on the other hand only tests 2 variables: Description, through descriptive and summative tasks, and Justification, through an argumentative task. Additionally, the TAST includes a task on following instructions, which is more suited to general language ability assessment, while a task on *giving* instructions would be a more useful task for testing the language of ITAs in a leadership position (see Table 4.2 for a summary of these findings).

Based on these findings, the comprehensive nature of the TBEST, and the deficiencies of the TAST in assessing the tasks that are a representative sample of the ITA instructional domain, I find that the TBEST has significantly greater domain validity.

5.1.2 Predictive criterion analysis. This research question requires that I search for predictive correlation between the TBEST and the TAST and student evaluations of the ITAs (using TCE scores). Unfortunately, there are no overt positive correlations. Also, the population of this correlation study is small (n=11).

The results of this analysis are tempered by two caveats. In order to conduct this predictive criterion correlation analysis, the population that I examined had to have taken both the TBEST and the TAST, and to have taught as the primary instructor (or in one case as a team-teacher) and to have been evaluated using the TCE. In the roster of 335 TBEST examinees there were 195 who had taken both exams. Of those 195, 122 earned a score that qualified them to be a primary instructor, but only 11 had TCE scores. This is a much smaller population than I anticipated, so the results should be cautiously received.

Secondly, none of the correlations between the four TCE questions and the TBEST or TAST scores are significant at either the .01 or .05 levels. Since this is an admittedly small population, with statistically insignificant results, only preliminary trends can be discussed. A larger comparison population needs to be analyzed to see if significant correlations exist.

With these caveats in mind, a correlation matrix (Table 4.4) reveals that there is a small correlation between the TBEST and TCE scores on questions of ITAs' overall effectiveness and the amount that students felt they learned (r=.237, .203, respectively) with a significance at approximately the .52 level.

A more perplexing result is found when looking at the TAST correlations. On each of the TCE questions, the TAST scores have a negative correlation, so higher TAST scores correlates with lower TCE scores. On Questions 1, 3, and 4 the correlations are (r=-.376, -.311, -.294) all at approximately a .33 level of significance. Question 2 is the most negatively correlated at (r=-.539) with a significance of .087.

Table 5.1

Correlations of TCE Results with TBEST and TAST (abbreviated version of Table 4.4)

		Q1*	Q2	Q3	Q4
TBEST	Pearson Correlation	.237	-.350	.203	-.147
	Sig. (2-tailed)	.482	.292	.549	.666
	N	11	11	11	11
TAST	Pearson Correlation	-.376	-.539	-.311	-.294
	Sig. (2-tailed)	.254	.087	.351	.381
	N	11	11	11	11

**All questions rated on 5-point Likert scale. 5=positive response, 1=negative response
Question 1: “What is your overall rating of this instructor’s teaching effectiveness?”
Question 2: “What is your overall rating of this course?”
Question 3: “How much do you feel you have learned in this course?”
Question 4: “What is your rating of this instructor compared with other instructors you have had?”

There are no discernable correlation patterns when data is re-organized by high/low average TCE results, or high/low TBEST or TAST scores. All correlations are essentially random. I agonized over what could cause the lack of correlations and here are my speculations:

a) TCEs do not measure language ability

b) TCEs do not measure teaching ability

c) Language ability doesn’t predict teaching ability

Each of these hypotheses are based on the premise that both the TBEST and TAST do measure language ability.

The first option is that the TCEs and the predictive assessment tools (TBEST and TAST) are measuring different variables, and this is why they have low correlation between their scores. The predictive tests are specifically language tests. The TCEs on the other hand measure a wide range of variables: course content, amount of time students spent studying in and out of class, teachers' respect of students, teaching ability relative to other teachers' ability, and other variables. Perhaps if more language specific questions were added to the TCEs there would be a higher correlation, at least to those questions. Possible questions could include: a) the instructor's rate of speech is understandable, b) the instructor's pronunciation does not interfere with my understanding of the course content, c) the instructor organizes lectures in a clear manner, d) the instructor checks for student comprehension during lectures and/or discussions, e) the instructor answers student questions in class or during office hours, and f) the instructor describes and explains course material effectively. These language specific questions could apply to ITAs as well as American TAs and professors as they highlight general communicative competence not specific to an English language learner.

The second option questions if the TCEs actually measure teaching ability. The TCEs are designed to be "teacher course evaluations" by definition, but perhaps the emphasis is more on the course variables, than on variables of teaching ability or effectiveness. There isn't broad consensus of what constitutes an effective teacher, the qualities that make some teachers effective might not be effective for other teachers or other disciplines. However, questions could be asked about classroom management variables, which could be relevant across multiple disciplines such as, a) instructor motivates students to pay attention in class, b) instructor conducts the class in an orderly fashion, c) instructor neutralizes classroom conflicts, and d) instructor addresses student questions or concerns in a timely manner.

The third option raises the question of language ability predicting teaching ability. This is clearly the predominant assumption because there is no teacher screening for proficient-English speaking TAs. Institutionally, there is an assumption that if you speak English well and are accepted for graduate study in your chosen field, that you are qualified to teach undergraduate students in your field. This language assumption gets challenged when ITAs are hired, and their only screening variable (beyond graduate school admission in their field) is their English ability. If and when students are dissatisfied with their undergraduate instruction by ITAs, after their language ability has been certified, there is no other variable to blame other than the language screening tool-even though an absence of teacher training might be at the root of the classroom dissatisfaction. So, perhaps the lack of correlations between the predictor evaluations and TCEs are simply a statistical demonstration that language ability does not predict teaching ability.

The third option is the likely culprit. While a hiring test is rightly needed to measure the speaking skills that are foundational to leading a class, actual success in the classroom involves much more than just speaking ability. The speaking hiring test should be seen as a tool to determine which candidates have the minimal language ability necessary to become an effective teacher, but then the department or the graduate college should take the responsibility of providing sufficient teacher and classroom management training to help ITAs actualize that potential.

5.2 Discriminative Utility

One route of analysis that I did not include in my initial research questions, but which I am considering post-hoc, is discriminative utility. Tests need to reveal differences in the assessment population as a function of their utility. In the case of this hiring assessment, it is necessary to separate the incoming ITA hiring pool into groups based on their level of instructional qualification, in the most reliable and efficient manner possible.

The pass rates listed in Table 5.2 reveal one of the greatest weaknesses of using the TAST as a hiring prerequisite; the huge score span of the “Fair” category (scores 18-25). Over 80% of the ITAs who took the TAST fall in the “Fair” category, requiring department interviews to confirm their eligibility to teach. Following this screening policy is no policy at all; there is no standardized minimum level of English proficiency when 84.4% of the hiring decisions are made on a case-by-case basis, and by the very departments who wish to hire them. The TAST cut score policy is then not reliable, since each department has a vested interest in positively assessing their own TAs, and it is not efficient since such a large percentage must undergo a secondary round of assessment.

Table 5.2

TAST and TBEST Pass Rates

TAST categories	**N**	**Result**	**Score Percentage out of N tested**
TAST score 26+	7	Qualified to teach	3.6
TAST 9-25 ("fair" 18-25) ("limited" 9-17)	186 (163) (23)	Interview by dept[9]	96.4 (84.4) (12.0)
Didn't take TAST	142	Take TAST, Interview by dept if <26	
TBEST categories	**N**	**Result**	
TBEST score 7 +	196[10]	Eligible to teach	58.5
TBEST score 6	117	Eligible to be lab/discussion leader, or to teach foreign language (ITAs' L1)	35.0
TBEST score 3-5	22	Not eligible, retake TBEST after 1 semester	6.5

Even if further cut-score analysis was conducted to pass a higher percentage of ITAs based on TAST scores, the issue of domain-validity remains. As I stated previously in section 2.2, all standard setting measures and cut-score analysis are worthless if the test content is not appropriately matched to the assessment goal, and the TAST content does not measure the ITA domain.

The TBEST pass rates, on the other hand, allow the graduate college to see the potential ITAs distributed across a standardized scale that denotes 3 levels of instructional eligibility. The TBEST cut-scores and test content resulted in 196 students

[9] Xi (2008) suggests that 24 should be the cut score for a "provisional pass" (required to take a training course while teaching their first semester). In this case, 24 more ITAs would be eligible to teach but 162 would still need to be interviewed or retake the exam.

[10] Quotas were not used to identify the number of TAs that needed to pass at the 6 and 7 levels; these are the results that occurred naturally. The cut scores were set by department representatives across the UA campus with the largest TA requirements. Each department has the discretion of setting individual cut scores, based on the criteria descriptions, but the general levels were agreed upon based on the description of the abilities described at each score level.

passing and eligible to teach, and 117 qualified to teach in a limited context, as lab or discussion leaders, or in a class teaching their first language. Just 22 of the examinees were deemed ineligible to serve as TAs at the time of their test taking.

These results show that the TBEST produces more reliable and efficient results, since there is not a reliance on a secondary screening, through departmental interviews. Additionally, the TBEST has greater domain validity than the TAST.

Table 5.3

Overall ITA Hiring Rates

Hiring periods	**Hired**	**Passed TBEST (6+)**	**Difference**
Fall 08	156	121	35
Spring 09	60	26	34
Fall 09	132	166	-34

Note: The number of ITAs hired includes both those newly hired (not renewed or current) ITAs who previously passed the hiring prerequisite, ITAs from Canada, the UK, or Australia who don't have to take an English ability test, and those who had a TOEFL 27+ who did not have to take the TBEST.

There is also a group of ITAs who were hired but who did not have to take the TBEST due to their TAST score of 27+ whose data is therefore not available for detailed analysis (Table 5.3). Between 35 and -34 ITAs were hired each semester over (or under) the number of ITAs who were approved to teach according to their TBEST results. It is not clear how many of these ITAs students are roll-overs, i.e. who passed the TBEST before there was a position available in their departments, or ITAs who were not required to take the TBEST. Regardless, they are a relatively small sub-group of the overall ITA population who have been deemed eligible to teach according to the TBEST.

5.3 Discussion of Over-arching ITA Issues

5.3.1 Training. Most ITAs that I surveyed view their experience as TAs as training for their future occupation. Administrators would be wise then to view ITAs as the future international faculty hiring pool, and provide comprehensive training to develop the pedagogical and communication strategies of ITAs. The desired content of ITA training received mixed results on the follow-up survey. By and large, ITAs want continued language (speaking and pronunciation) instruction; even after their language ability has been evaluated and deemed proficient enough for instructional responsibilities. The assumption about language ability predicting teaching ability seems to be prevalent even among ITAs. Most ITAs do not have pedagogy training or teaching experience prior to serving as ITAs, and despite acknowledging pedagogical difficulties in their classes, they still do not prefer to be taught teaching methods in an ITA training course. I think administrators need to do what is best for ITAs, not just what is popular in this case. Clearly training is needed, and pronunciation drills won't produce the desired pedagogical improvements.

5.3.2 Evaluation. Most ITAs do not get evaluated on a regular basis because they are not usually assigned to positions as primary instructor. ITAs are missing the opportunity to learn through periodic performance review. Even though performance evaluation is intimidating, it is something that they will have to do on a regular basis if they decide to pursue careers as professors.

Furthermore, ITAs need to find ways to use student feedback to make mid-term adaptations. When student feedback is primarily used as an effectiveness metric, and only after the semester is over, it creates a distorted power structure, where students have power over ITAs, who then feel like they can't give justified criticism of student work for fear of negative student evaluations. With more of an emphasis on mid-term formative evaluation, students and ITAs can become more cooperative members in the success of classroom instruction, because students assume some of the responsibility of identifying variables for improvement.

5.3.3 World Englishes. World Englishes are the (Indian) elephant in the room when it comes to ITA hiring standards. What should be done with ITAs who speak English as their first language, but their English variety happens to have a dispreferred accent? With the right social context, listeners will adapt and compensate for a wide variety of accents, and likewise, if they feel justified in their accent prejudice, listeners will obstinately refuse comprehension. Just because an accent is not preferred should not be grounds for excluding an applicant.

For most ITAs, their language development and pronunciation is probably fossilized at this stage. To be studying at the graduate level in a second language, they have already finely-honed the language skills necessary for that level of communication, and developed the compensation strategies that they personally need. Most likely they are not still active language learners and so they should embrace their language as it is and instead focus on perfecting the performative skills that their position requires: being an engaging public speaker, learning strong classroom management skills, writing academic articles, or presenting scientific research. As their proficiency in these skills improves, language fluency will likely improve as a positive side effect of that increase in confidence and extended practice.

Undergraduate students need to experience diversity and take more personal responsibility for their education. As one of my survey respondents said, “accents are never a problem for good students, because those students are actively involved in learning, asking questions, doing the homework, coming to office hours and they choose to learn.” Accents are more likely to affect the students who are looking for something to blame other than their own dedication to the subject material or involvement in class. With supportive pedagogy strategies, ITAs can help students adapt to different accents and understand the course content.

A second area of World English concern is looking at ways that the TBEST can be more accessible to ELF (English as a Lingua Franca) speakers. Elder and Davies (2006) outline five practical accommodations that can be made:

1. Vet texts for topics or genre that may exclude non-standard English users
2. Define or avoid lexical items that would be unfamiliar to NNESs
3. Use ELF interlocutors
4. Train raters to overlook errors that do not result in miscommunication
5. Involve ELF users in standard setting exercises

(Elder & Davies, 2006, pp. 289-290)

Currently, the TBEST development team already employs the first two suggestions in the development process of selecting prompts and crafting question instructions. The next step should be implementing points three and four. It would be beneficial to employ ELF interlocutors to serve as TBEST administrators. Through this step, ELF examinees will be made more comfortable and will be confident that ELF speech is valued. Furthermore, while the current rubrics do not uphold 'native-speaker' norms, the rating criteria could be changed to lessen the weight of the Accuracy and Fluency criteria, and place extra weight in the Overall Message and Complexity categories to demonstrate that communicative competence is most valued in this assessment. The fifth suggestion could be implemented at a future date, after more data is collected. Since this is a hiring tool that is implemented across more than 50 departments, it is important that the departments believe the standards are justified and useful, and having actual data samples of ELF examinees may make the standard setting process more concrete and less theoretical.

5.3.4 Departmental misuse. Some ITAs expressed frustration that hiring policies were not transparently stated or uniformly enforced. I heard difficult stories of ITAs teaching before taking the hiring test, and then losing their positions mid-semester, due to incorrect application of cut scores. Other students were transitioned back and forth between RA and TA positions without being tested and meeting the language requirements. These students were not paid while waiting on professors' grants to come through, and felt powerless to object, since they seemed to be operating outside the official system.

There seems to be some neglect of ITAs' circumstances in some departments. Moving to a new country for education and employment is stressful and very expensive. Hiring departments need to do what is possible to ensure ITA funding before the ITA arrives. For tests like the TBEST, this means that procedures need to be developed so that graduate students can take the test while in their home country so that they know what their funding options will be in their first year. The biggest hurdle is security, verifying the identity of the test taker. However, in every high-stakes testing scenario some people will be motivated to try to cheat the system, there will just have to be vigilance of verifying the results once the ITAs are at the university.

To further equip ITAs, I advocate publishing ITA hiring and employment guidelines online so that ITAs know what are the reasonable expectations for their positions and to identify a reporting procedure if their department is in violation. Care should be taken to make sure that ITAs do not face departmental retribution for reporting violations.

5.3.5 ITA excellence. ITAs had comparable and even higher TCE score means than the university provided comparison group. Perhaps this is due to the population surveyed. Eight (out of eleven) were foreign language instructors, so their status as a native speaker of that language would actually make them *more* qualified in the eyes of their students than an American TA. Two taught engineering courses, where presumably, there is a positive bias that other countries have better math and science education than in the U.S., so again, an international TA might have more credibility than an American TA.

None of those ITAs sampled for TCE analysis were in TA positions that would cause students to directly challenge the ITAs' role as experts, such as ITAs teaching English Composition. Several Chinese ITAs who teach English composition stated in personal conversations that they must constantly establish their authority in the classroom because students feel entitled to challenge their teaching and grading on a regular basis.

5.4 Implications for ITAs and Future Studies

5.4.1 ITA hiring procedures. As I began surveying the policies at schools across the country I saw a need for this type of survey to be published. A valuable follow-up to this dissertation would be to expand my survey of top research and teaching universities, and to publish the overall hiring, training, and evaluating policies at these universities. I think that schools implement ITA policies out of a need to prove they are doing *something* in case students or parents complain about ITAs, but not necessarily because those policies are the most effective for preparing ITAs to teach undergraduates. With a comprehensive survey of current policies, administrators could easily look at what other peer institutions are doing, and use that information for improving their own policies.

Virtually every school has testing and screening procedures for ITAs, but mandatory training programs were sparse with an emphasis on efficiency, focusing on the minimum acceptable training, rather than on what ITAs really need. Finally, evaluation programs were incredibly difficult to find. I only found one school that had extensive evaluation protocols in place to provide ITAs with *formative* feedback, rather than just a summative pass/fail type evaluation report to put in the ITAs' teaching files. An overall survey could help universities balance out the ITA circle of learning through these three steps.

Administrators and hiring departments need to see the ITA experience as a complete cycle that is just beginning when the ITA passes the initial screening exam. The chance to get teaching experience while pursuing graduate studies is one of the key differences between the American and British graduate systems, and ITAs who choose to come to the U.S. probably considered that variable in making their decision. Therefore, to make the TA opportunity the most valuable for incoming students, it should be more than trial-by-fire; there needs to be a pedagogical support system in place to guide ITAs, especially in their first years of teaching.

One specific improvement that I hope will result from this study is that an emphasis be placed on teacher training rather than always emphasizing accent elimination, or other such impossible tasks. I think it is an unfortunate side effect of the monolingual bias in the United States that students only value native-like proficiency. As it is, many students seem personally intimidated by learning foreign languages, and assume that they have to speak *perfectly* in order for their learned language to be useful. Then this expectation of perfection gets transferred to ITAs. In fact, ITAs with strong interpersonal skills, effective classroom management, and sufficient content knowledge can easily compensate for distinctly non-native English ability. The native-speaker ideal is unrealistic and an impractical expectation to hold ITAs to.

In a recent New York Times article, former teacher, principal, and current educational consultant, Doug Lemov discusses the results of a five-year study of effective teachers. Teachers were initially selected when they were able to "squeeze high performance out of the poorest students" (Green, 2010, p. 7) and then Lemov interviewed them and videotaped their classrooms. Lemov found that classroom management, being able to capture students' attention and get them to follow instructions, was the common denominator among these successful teachers. Lemov compiled the best 49 teaching techniques into a book, which is now being implemented at school districts across the country, in immersive teacher training programs, and by Teach for America. Even though this resource is intended for the K-12 instructor, it still argues an important principle: "great teachers are not born, but made" (Green, 2010, p.15) and it could empower ITAs to make concrete improvements to their teaching effectiveness instead of having a deficit mentality that forever focuses on their lack of native English ability. This could be a great resource for ITAs because it puts the focus back on the classroom, and not simply on the foreign-accent scapegoat.

Three specific training techniques that ITAs should be taught are: the importance of establishing a classroom routine, methods for eliciting participation during discussions, and repetition strategies. First, many ITAs struggle with classroom management and having a routine gives the ITAs a way to establish authority and also guides lesson preparation. Students also learn better when information is structured and predictable, so this may help lessen communicative anxiety that they may have about interacting with an ITA. Second, the interactive classroom presents one of the most daunting cultural challenges to many ITAs, and hosting a discussion course where no one talks is any teacher's nightmare. ITAs should be explicitly taught methods for encouraging group and class participation. Third, students have diverse learning preferences, but it is easy for teachers to rely most heavily on the strategy that they personally prefer. By teaching creative ways to repeat and practice the same or similar information and skills more students will benefit from the ITAs teaching. While I have made the argument that these teaching strategies should be taught to ITAs due to the scope of this dissertation, the recommendations are universally valid. All graduate TAs would benefit from this training.

5.4.2 Future studies. This study has broad applications for assessment theory in general, and for improving the TBEST and ITA language assessment specifically. Future studies could also focus on World English data within this assessment data set and in the development and analysis of ITA training programs and evaluating protocols.

I think that this study did many things well, primarily because I had the support of the Graduate College and therefore had access to the entire population of incoming ITAs at the UA. Over 30 countries and 50 departments are represented by the students in this survey. Through this diversity I am able to avoid disciplinary and linguistic idiosyncrasies that could skew the overall data. I was pleased with the diversity of the ITA population that were tested, and with the strong response to the follow-up survey. The follow-up survey had a 35% response rate (119 of 335), which demonstrates that ITAs care about improving the ITA process and were eager to share their impressions.

One major weakness was the lack of criterion data for my second research question. Since I formulated my research questions while still collecting data, I didn't know what would be available to analyze and I assumed that the majority of ITAs would receive the TCE scores. I never would have guessed that only 11 TCE scores would be found[11]. Part of my miscalculation comes from my own disciplinary bias. In the English department where I served as a TA, there were no lab or discussion sections, all TAs were the primary instructor for the class that they taught. I mistakenly thought that was the norm. According to my survey, most other departments use ITAs much less frequently (just 20%) as primary instructors (see Table 4.5 for complete results).

5.4.2.1 Assessment theory. This study contributes to assessment theory in general by serving as a wake-up call to test developers and administrators about the importance of respecting an assessment tool's intended use. There needs to be a return to prioritizing the *validity* of a particular assessment tool. Throughout presentations on this TBEST study I have often been asked, What is the correlation of TBEST to the TOEFL? and What scores are equivalents? While there is high correlation (r=.592), and though it is possible to find TOEFL scores at which it is likely that those candidates will also score passing TBEST scores, the danger is that for the sake of efficiency, just the TOEFL score would then be used, instead of relying on the test that is actually suited to the assessment need at hand, determining teaching qualification, as opposed to simply admittance qualification on the TOEFL.

5.4.2.2 Improving the TBEST. There is still a lot of data left unanalyzed in this body of research. Specifically, the TBEST could be re-analyzed through inter-item analysis; perhaps not all of the questions are necessary and the TBEST could be a little shorter. Question 2, for example, (the interpretation of a proverb) was a question that

[11] The population that I searched for TCE scores was two-thirds of the total TBEST examinee population, only those ITAs who also took the TAST were considered since I needed to be able to compare both predictor scores to the TCE criterion results.

some ITAs grumbled about, not seeing the direct application to the classroom. Although it has value for the reasons described in the Analysis chapter, I would be curious to see how the overall scores changed if that question were omitted.

I would also recommend that the TBEST be made even more tailored to the teaching context and that tasks that overlap with the TAST be removed. Both tests have Descriptive tasks, perhaps that is not necessary on the TBEST, or perhaps it is used simply as a warm-up and not rated. The most unique test content on the TBEST is the video of teaching scenarios, and adding another video task could strengthen the TBEST validity even more. Since the majority of ITAs do not teach classes individually, but work more closely with students as discussion or lab leaders, I think there should be an Advising task that is placed in an office-hour context. By making the TBEST content even more specialized to the ITA context, the test scores on TBEST and TAST will hopefully be less likely to be conflated or used to serve both hiring and admittance purposes.

To further improve the TBEST, I think considerations should be made to change the reporting format in two ways: to report task scores individually, and to expand rating bands. Currently only the overall score, the average of the four tasks, is reported to departments and to students. However, it may be beneficial to departments to see particular scores on individual tasks as they would pertain to the specific ITA position that they must fill. An ITA position as a primary instructor might lead departments to focus on the score on the Interpretive and Comprehensive Video tasks, while for a discussion or grading position they might prioritize the scores on the Descriptive and Argumentative tasks.

Secondly, the reporting system that is currently used only allows scores to be reported as integers, so students scoring 6.01 and 6.83 are both reported as a 6. In this way there is no distinction made between the low 6 and the high 6 and departments may be willing to hire ITAs with borderline scores. It would be more beneficial for the rating reports to be expanded to include either 2 or 3 sub categories (high/low or high/mid/low) within each of the score bands. This is the opposite problem than what exists on the TOEFL where there is a huge band for the "fair" category for example, from scores 18-25, and all of these students would have had to be interviewed in a secondary screening in order to qualify them as ITAs at the UA. The overarching problem with bands is that they can conflate the abilities of all those who score within the given band if the band itself is not further defined. An example of implementing a simplified scale, which still includes basic intra-band description, is the Common European Framework (CEF). The CEF has 3 main bands: Proficient, Independent, and Basic User, and then each of those bands is subdivided into high and low divisions (Common European Framework, 2001, p. 23).

5.4.2.3 World English accommodations. World Englishes could also be studied in future research. Demographic information isn't currently assigned to the individual identification numbers or sound files, but it would be possible to coordinate with the Graduate College to collect that information and merge it with the test identification numbers that were used. It would be interesting to see if there are trends of success or difficulty, by task, or by criteria category (overall, accuracy, fluency, organization) based on the ITAs' country of origin. This study could indicate if accommodations should be made to facilitate World English speaking examinees.

5.5 Reflections

This dissertation project was a valuable learning experience for me, especially the process of creating an assessment tool. I discovered how frustrating it is to try to create a fair, valid test, and then learn that it is not being applied as intended. Our development committee was very deliberate in the process of trying to make each decision purposefully, not just out of convenience. We discussed each prompt, the wording for every question, the terms to be used in the rubric, every facet.

My first disappointment came when I looked at the first round of test results and suspected that the raters were holistically rating the speech samples instead of analytically rating the separate criterion individually. I was initially frustrated that the raters didn't follow our instructions and were instead relying on gut-instincts, or at least that was my impression. I was able to step back and realize that the provided rubric was not as useful as it could be; after rewording some easily conflated categories and retraining the raters, this problem was resolved.

However, more frustration came in the misapplication of the test by individual hiring departments. From my perspective, the departments are taking advantage of, or simply disregarding the needs of ITAs by requiring them to retake the test multiple times when the general cut score had already been achieved. Of course, I cannot presume the actual department rationale behind their actions, perhaps they have honestly mistaken the cut score recommendations, or have set higher requirements for their department.

With this lesson in mind I have some sympathy for the TOEFL creators. Their test was never intended to be used to screen ITAs; it is purposefully a generic, general English ability assessment tool to be used to inform admission decisions for English medium universities (Xi, 2008). However, when it became apparent that many schools were misapplying the TOEFL-iBT TAST section for ITA screening, ETS' response was inappropriate. Instead of creating a tool that would be appropriate for this population, as a potential sub-test, they instead convened a simple ITA standards setting committee to set recommended cut scores for hiring ITAs (Wylie & Tannenbaum, 2006). So, even though they knew that their TAST *content* was insufficient to measure ITAs' speaking ability, they validated this practice by setting recommended TAST cut-scores. A possible solution would be to incorporate the TBEST as a sub-test specifically for screening potential ITAs.

5.6 Conclusion

The solution to complaints against ITAs is often thought to be stricter language requirements, not a greater emphasis on mandatory teacher training. It is absolutely necessary require minimum English speaking ability of ITAs, but there must also be pedagogical support to enable them to be successful teachers. Language testing and pronunciation drills alone will not raise the level of ITA effectiveness.

The ITA experience has great potential to transform the education of both international graduate students, and the general American undergraduate student body.

For international students, teaching an undergraduate course may be their first opportunity to see unscripted American culture lived out, without the saccharine lens of Hollywood, the perfection of a guided tour, or the complications of inter-departmental hierarchy and political posturing. The comments I received on the follow-up survey demonstrate this opportunity:

- *Students I've taught were actually great. They're fun and I actually learned a lot from them as well, e.g. how American students solve problems, how they interact w/ each other, how slangs actually work, etc.*
- *It's quite an amazing experience to be a TA.*
- *Being a TA helps [you] to learn how to speak up; to develop self-esteem and, to improve the language skills*.

Even with the difficulty of balancing the commitments of both teaching and taking classes, the results of the survey reflected an ITA population that is overall satisfied with their TA position.

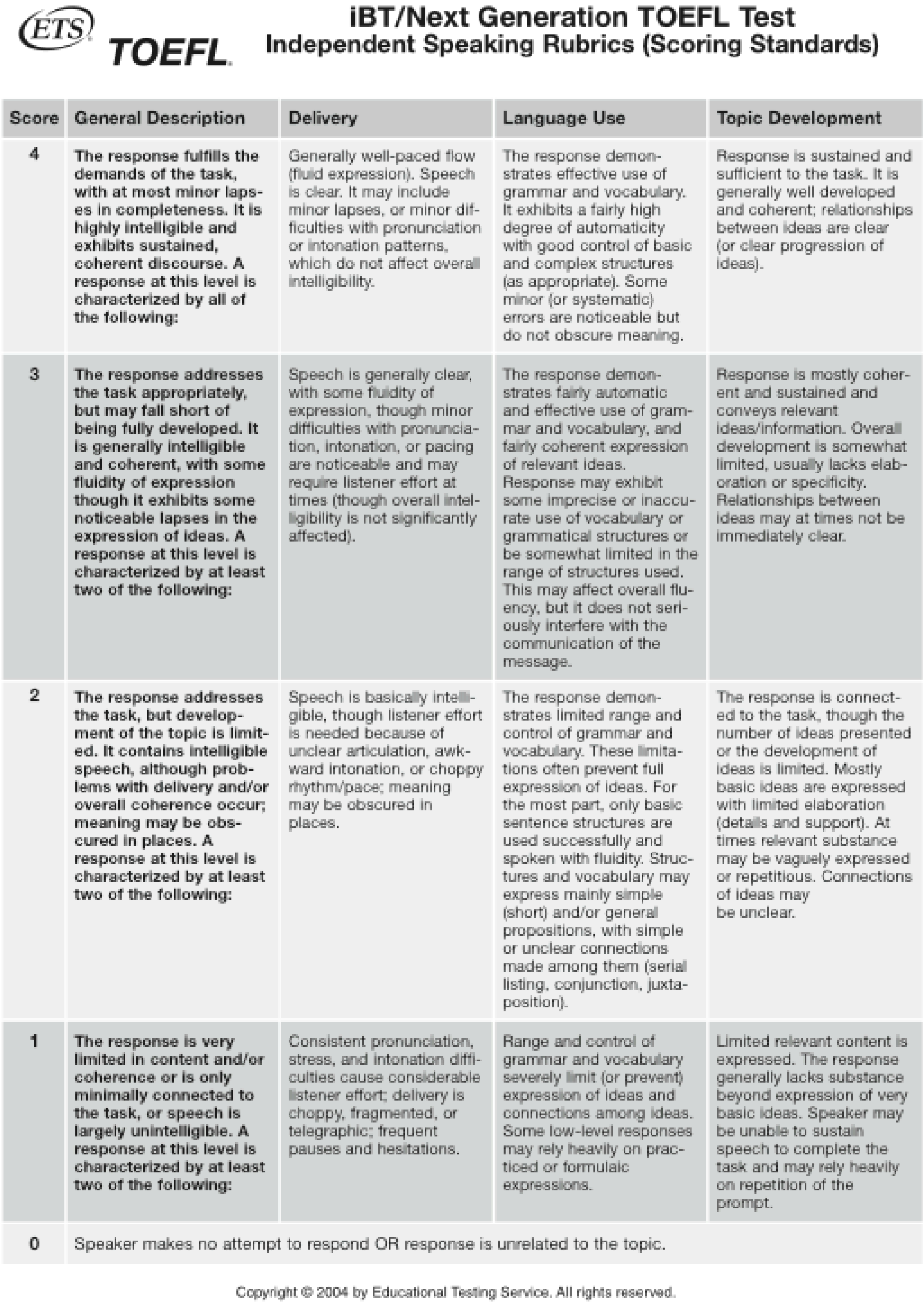
For undergraduates, having an ITA gives them the opportunity to experience new points of view and practice global communication skills. Virtually every sector of American society is becoming more globally inter-dependant, and students need to develop a higher tolerance of difference and ambiguity. The diversity represented in education is representative of the diversity that students will have to work with in the workplace. Successful students will be those who are willing to step outside their comfort zone and learn how to cooperatively adapt to new communication strategies, speaking styles, and learning patterns, not those who simply expect everyone to adapt to them.

The ITA employment process can be complicated for administrators, but with greater transparency between institutions, there can be an exchange of best practices. As the process becomes smoother it will be easier to focus on all the benefits of a robust ITA program, bringing diversity to graduate programs, equipping the next generation of international faculty, and preparing undergraduates to excel in a globally diverse society.

APPENDIX A:
TBEST Assessment Rubric: *Advanced & Intermediate samples*

Band	Overall Performance	Accuracy (Word Choice & Grammar)	Fluency (Pronunciation & Intonation)	Complexity (Organization)
9	The task is done very competently and with ease. The quality of speech is polished and professional, although not at every moment. When difficulties in expression are encountered by the speaker, he or she quickly chooses an effective way to rephrase the message.	Speakers at this level have strong control over English grammar, although errors do still occasionally occur. Word choice is varied and complex	Speech is produced easily and with confidence most of the time. Only in the case of specialized topics or in the midst of a rather abstract argument or discourse structure will the speaker run into some problems, but the speaker is adept at circumlocution and other types of restating that quickly get back into the flow of what was being said. Pronunciation is nonnative-like but does not impede comprehension.	Answer is coherent, with good transitions and conjunctions. Sentence structure is quite varied and mostly complex, with just a few simpler sentence forms. No reformulations are necessary.
7	The requirements of the task are fully met, although the answer still lacks the fullness of speech of someone used to living and dealing with other people in that language on a day-to-day basis.	The speaker's use of vocabulary is adequate but generic. The intended meaning is generally clear, but sometimes the words chosen are not quite accurate. The choice of grammatical structures is at times distinctly nonnative-like.	The speaker can speak smoothly most of the time. Occasional difficulties, however, result in increased hesitance, shortened and/or choppier sentences, and increased self-correction. Smoother speech is associated with everyday topics, whereas roughness comes about with more advanced topics. Pronunciation is non-native but doesn't often impede comprehension.	Topic is logically organized: speaker uses transitions between points. A variety of sentence structures and transitions are used. Doesn't rely on repetitive and simple sentence structures. While quite adequate, the complexity is clearly not of the sort that would be used naturally by someone speaking in his or her first language.

APPENDIX B:
TAST Assessment Rubric



iBT/Next Generation TOEFL Test
Independent Speaking Rubrics (Scoring Standards)

Score	General Description	Delivery	Language Use	Topic Development
4	**The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:**	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	**The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:**	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	**The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:**	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	**The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:**	Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit (or prevent) expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

APPENDIX C:
International Teaching Assistant Survey

Thank you for participating in this brief survey.
Your answers will NOT be reported to your department and will have NO effect on your hiring status. This survey is intended to collect information to make the ITA hiring, training, and evaluating process better for future ITAs.

The questions in this section give us demographic and course information about international teaching assistants at the University of Arizona.

1. Number of years of teaching experience do you have?
None
1-3
3-10
10+

2. Where have you taught?
A school in your home country (any age level)
University of Arizona
Pima Community College
Another U.S. university
Other
I haven't taught

3. Are you eligible to serve as a TA?
Yes, I passed the speaking test (TBEST)
No, my score was too low for my department, so I will not be a TA
No, my score was too low, so I will re-take the TBEST

4. If yes, describe your current/most recent TA position:
Class size:
Sm: <20
Med: 20-40
Lg: 40-100
Ex Lg: 100+

Primary Your role:
instructor
Lab Leader
Discussion Leader
Team Teacher
Grader
Individual Tutor
Other:

5. What types of speaking tasks does your TA position require? Rank the frequency of the following: 9=most frequent, 1=least frequent

Explaining /Elaborating
Describing something
Giving suggestions/advice
Reading aloud from a text
Justifying grades/homework points
Restating information in a simple way
Giving step-by-step instructions
Disciplining students
Reading aloud from a book
Giving formal presentations
Comments:

6. What parts of teaching at the UA were difficult for you?
Rate the following 5= most difficult, 1= least difficult

Leading student discussions
Presenting lessons in class
Knowing when students understood/were confused
Planning lessons
Grading
Creating assignments
Creating tests
Explaining assignments to students
Solving conflicts between students
Answering student questions in office hours
Explaining grades to students
Dealing with cultural differences with students
Using D2L (online course management program)
Using classroom technology
Other:

The questions in this section gives us input about improving the training available to international teaching assistants at the University of Arizona.

7. Would you be interested in a TA training course?

No, TA orientation is enough
Yes, a 1-week training course
Yes, a semester-long training course
Yes, a year-long training course

8. When would you prefer to take the training?

Before taking the speaking test (TBEST)
After taking the speaking test, but in the semester prior to being a TA
During the first semester of being a TA (concurrently)
After the first semester of being a TA
Comments:

8. What would you like the training course to include?
Rate the following, 5=most useful, 1=least useful

General teaching strategies
Practice teaching scenarios with your class material
Cultural teaching tips
Presentation skills
Pronunciation practice
Developing syllabi/assignment sheets
Speaking fluency practice
Powerpoint demonstrations
Faculty mentor meetings
Having your classroom teaching observed
Observing other classes
Other:

The questions in this section will give us input about improving the teaching evaluation available to international teaching assistants at the University of Arizona.

Evaluating Questions

9. When students rate ITAs what do you think students value most?
Rate the following, 5=most valued, 1=least valued

ITA's Content knowledge
Easy Grading
Clear pronunciation
Feeling respected by the ITA
Fairness in Grading
Humor
Feeling comfortable asking ITA questions
Other:

10. How would you like to be evaluated?

Rate the following, 5= most desirable, 1= least desirable.
Student end-of-term evaluations (you see feedback after turning in final grades)
Student mid-term evaluations (you see feedback right away)
Advisor observation
Self-evaluation
Evaluation of materials you designed
Evaluation of student progress (student passing rate)
Peer-evaluation (another ITA observes your class and evaluates)
Other:

11. Overall, how do you feel about being a TA?
Rate the following, 5=strongly agree, 1= strongly disagree.

It is good training for my future
It saves me a lot of money
Students respect me
Students treat me the same as an American TA
It will make me take longer to graduate
If I wasn't a TA, I would learn more in my own classes
It doesn't take very much time to be a TA

12. Do you want to tell us anything else about being an ITA?
Write your comments here!

REFERENCES

Anastasi, A. & Urbina, S. (1997). *Psychological Testing*, 7th Ed. Upper Saddle River, NJ: Prentice Hall.

Bailey, K. (1983). Foreign teaching assistants at U.S. universities: Problems in interaction and communication. *TESOL Quarterly, 17*(2), 308-310.

Bailey, K. (1984a). A typology of teaching assistants. In K. Bailey, F. Pialorsi, & J. Zukowski-Faust (Eds.), *Foreign teaching assistants in U.S. Universities (pp. 110-130)*. Washington, DC: National Association for Foreign Student Affairs.

Bauer G. (1991). Instructional communication concerns of international teaching assistants: A qualitative analysis. In J. Nyquist, R. Abbot, D. Wulff, & J Sprague (Eds.), *Preparing the professoriate of tomorrow to teach (pp. 420-426).* Dubuque, IA: Kendall/Hunt Publishing.

Black, P. (2009). Formative assessment issues across the curriculum: The theory and practice. *TESOL Quarterly, 43*(3), 519-523.

Brewer, J. & A. Hunter. (1989). *Multimethod research: A synthesis of styles.* Newbury Park, CA: Sage.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Bygate, M. P. Skehan, & M. Swain. (Eds.). (2001). *Researching pedagogic tasks: Second language learning, teaching, and testing.* Harlow: Longman.

"Common European Framework of Reference for Languages." (2001). Retrieved from http://www.coe.int/T/DG4/Linguistic/Source/Framework_EN.pdf.

"Conditions of Appointment for TAs who are not Native Speakers of English" at the University of Washington [Online policy website]. Retrieved from http://www.grad.washington.edu/policies/memoranda/memo15.shtml.

"Consulting FAQ" for ITAs at the University of Washington [Online policy website]. Retrieved from http://depts.washington.edu/cidrweb/consulting/ita.html.

Crookes, G. (1986). *Task classification: A cross disciplinary review.* Technical Report No. 4: Center for second language classroom research. Honolulu: University of Hawaii.

Davis, B. (1987). The effectiveness of videotaped protocols as a training technique for international TAs. In Chism, N. & S. B. Warner (Eds.) *Institutional responsibilities and responses in the employment and education of teaching assistants* (pp. 321-333). Columbus, OH: Center for Teaching Excellence.

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

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods.* NY: McGraw-Hill.

Dick, R. C., & B. M. Robinson. (1993). *Oral English proficiency requirements for ITAs in U.S. colleges and universities: An issue in speech communication.* Paper presented at the World Communication Association biennial convention, Pretoria, South Africa. Retrieved from ERIC database. (ED360653)

Douglas, D. & L. Selinker. (1989). Markedness in discourse domains: Native and non-native teaching assistants. *Papers in applied linguistics, 13*(1), 69-81.

Dunn, T. G & J.C. Constantinides. (1991). Standardized test scores and placement of international teaching assistants. In J. Nyquist, R. Abbot, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 414-419). Dubuque, IA: Kendall/Hunt Publishing.

Ecclestone, K., & Pryor, J. (2003). "Learning careers" or "Assessment careers"? The impact of assessment systems on learning. *British educational research journal, 29*(4), 471-488.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford UP.

Faggen, J. (1994). *Setting standards for constructed response tests*: *An overview* (ETS RM-94- 19). Princeton, NJ: ETS.

Falsgraf, C. (2009). The ecology of assessment. *Language Teaching, 42*(4), 491-503.

Finder, A. (2005, June 24). Unclear on American campus: What the foreign teacher said. *New York Times.* Retrieved from http://www.nytimes.com

Fitch, F., & S. E. Morgan. (2003). "Not a lick of English": Constructions of ITA identity through student narratives. *Communication Education, 52*(3/4), 297-310.

Freeman, G. (1996). *Judging oral proficiency: Can the native judge determine standardized test scores (TSE) through an interview process*. Dissertation, The Florida state university, School of Education.

Gallego, J. C. (1990). The intelligibility of three nonnative English-speaking teaching assistants: An analysis of student-reported communication breakdowns. *Issues in applied linguistics, 1*(2), 219-237.

Gareis, E. and Williams, L. (2004). International faculty development for full-time and adjunct faculty: A program description. *The journal of faculty development, 20*(1), 45-53.

Gonzlez Pino, P. (1988). *Testing second language speaking: Practical approaches to oral testing in large classes.* Newsletter. Middlebury, VT: Northeast Conference of language Teaching.

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

"General TCE Information" at the University of Arizona (Online policy website). Retrieved from http://aer.arizona.edu/teaching/Guide/TCEGuide.asp.

Green, E. (2010, March 7). Building a Better Teacher. *New York Times.* Retrieved at http://www.nytimes.com.

Hagstrom, F. (2006). Formative learning and assessment. *Communication disorders quarterly, 28*(1), 24-36.

Hendel, D. D., Dunham, T., Smith, J., Solbert, J., Tzenis, C., Carrier, C., and K. Smith. (1993). Implications of student evaluations of teaching for ITA development. In K.G. Lewis (Ed.), *The TA experience: Preparing for multiple roles* (pp. 390-400). Stillwater, OK: New forums press.

Hoekje, B, & K. Linnell. (Spring 1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly,* *28*(1), 103-126.

Hoekje, B & J. Williams. (1994). Communicative competence as a theoretical framework for ITA education. In C.G. Madden & C.L. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp.11-26). Alexandria, VA: TESOL.

Inglis, M. (1993). The communicator style measure applied to nonnative speaking teaching assistants. *International journal of intercultural relations 17*(1), 89-105.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes, 64*(4), 555-580.

"ITA Training Program" at the University of Delaware [Online policy website]. Retrieved from http://www.udel.edu/eli/ita2/index.html.

"ITAP: International Teaching Assistant Program" at Cornell University [Online policy website]. Retrieved from http://cte.cornell.edu/campus/itadp/courses.html.

Jacobs, L. C. & C. B. Friedman. (1988). Student achievement under foreign teaching associates compared with native teaching associates. *Journal of higher education, 59*(5), 521-563.

Jenkins, S., & D. L. Rubin. (1993). International teaching assistants and minority students: The two sides of cultural diversity in American higher education. *The journal of graduate teaching assistant development, 1*(1), 17-24.

Kulik, J.A., Chen-Lin, C. , Cole, M. A., & S. L. Briggs. (1985). *Student evaluations of* *foreign teaching assistants: Internal report.* University of Michigan: Center for Research on Learning and Teaching.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*(2), 151-172.

Lee, Y. (2000). Effects of degrees of task complexity on L2 production. In S. Kang (Ed.) *Korean language in America 5.* Monterey, CA: The American Association of Teachers of Korean.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166.

Lemov, D. (2010). *Teach like a champion: 49 techniques that put students on the path to college.* San Francisco*:* Jossey-Bass.

Long, M. (1985). A role for instruction in second language acquisition: Task based language teaching. In K. Hyltenstam and M. Pienemann (Eds.), *Modeling and assessing second language acquisition.* Clevedon: Multilingual Matters.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*(4), 415-437.

Lewis, K. G. (1997) *Training Focused on Postgraduate Teaching Assistants: the North American Model*, Retrieved from http://www.ntlf.com/html/lib/bib/lewis.htm.

Lewkowicz, J.A. (1997). The integrated testing of a second language. In Clapham, C. & Corson, D., (Eds.) *Encyclopedia of language and education.* 7, (pp. 121-130). Dordrecht: Klewer Academic, Liu, J. (1999). From their own perspectives: The impact of non-native ESL professionals on their students. In Braine, G. (Ed.), *Non-native educators in English language teaching,* (pp. 159-177). Mahway, NJ: Lawrence Erlbaum.

Messic, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement*, 3rd Ed, (pp. 13-103). New York: American Council on Education/Macmillan.

Mestenhauser, J.A. (1981). Foreign students as teachers: Lessons from the program in learning with foreign students. In G. Althen (Ed.), *Learning across cultures* (pp. 143-149). Washington: D.C.: NAFSA.

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

Myers, C. L. & B. Plakans. (1991). Under controlled conditions: The ITA as laboratory assistant. In J. Nyquist, R. Abbot, D. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach (pp. 368-374).* Dubuque, IA: Kendall/Hunt Publishing.

Nitko, A. J., (1989). Designing tests that are integrated with instruction. In R.L. Linn (Ed.) *Educational measurement*, 3rd Ed, (pp. 447-474). New York: American Council on Education/Macmillan.

Norris, T. (1991). Nonnative English-speaking teaching assistants and student performance. *Research in higher education, 32*(4), 433-448.

Norris, J. M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum, 39*(1). Retrieved from http://exchanges.state.gov/forum/vols/vol38/no1/p18.htm.

Numrich, C. (1993). Changing (and unchanging) attitudes and values of new ITAs: Training curricula implications. In Lewis, K. (Ed.) *The TA experience: Preparing for multiple roles (Selected readings from the third national conference on the training and employment of graduate teaching assistants),* (pp. 359-367). Stillwater, OK: New Forums Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge UP.

"Open Doors." (2008). (Online international student report). Retrieved from http://opendoors.iienetwork.org/?p=150813.

OIRPS. (2009). (University of Arizona Office of Instructional Reporting and Planning Support Website). Retrieved from www.oirps.arizona.edu.

Perie, M., Marion, S., & B. Gong. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessment. *Educational measurement: Issues and practice 28*(3), 5-13.

Prabhu N.S. (1987). *Second language pedagogy.* Oxford: Oxford UP.

"Procedures for the Evaluation and Certification of the English Fluency of Undergraduate instructional personnel." (Online policy website at the University of Pennsylvania) Retrieved from http://www.vpul.upenn.edu/osl/fluency.html.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science 28*, 4-13.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9,* 109-121.

Reinhardt, J. (2007). *Directives usage by ITAs: An applied learner corpus analysis.* Dissertation, Pennsylvania State University.

Reynolds, C. Livingston, R., & V. Wilson, (2006). *Measurement and assessment in education.* Boston, MA: Pearson.

Ronkowski, S. (1987). International and American TAs: Similarities and differences. In Chism, N & S. B. Warner (Eds.) *Institutional responsibilities and responses in the employment and education of teaching assistants* (pp. 263-266.). Columbus, OH: Center for teaching excellence.

Ross, S. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics, 26*(3), 317-342.

Rounds, P. L. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly, 21*(4), 643-671.

Rubin, D. L. (1992). Non-language factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in higher education, 33*(4), 511-531.

Saif. S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian journal of applied linguistics, 5*(1-2), 145-166.

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing, 23*(1), 1-34.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.

Scriven, M. (1967). The methodology of evaluation. In Tyler, R. W., Gagne, R. M. & M Scriven (Eds.) *Perspectives on curriculum,* (pp.39-83). Chicago, IL: Rand McNally.

Shepard, L. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational measurement: Issues and practice, 28*(3), 32-37.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*, 38-62.

Skehan, P. (1998). Task-based instruction. *Annual review of applied linguistics, 18,* 268-286.

Smith, R. M., Byrd, P., Nelson, G., Barrett, R., & J. Constantinides. (1992). *Crossing pedagogical oceans: International teaching assistants in U.S undergraduate education*. Washington, DC: The George Washington University. Retrieved from ERIC database. (ED358812)

Smith, K. S. (1993). A case study on the successful development of an international teaching assistant. *Innovative higher education, 17*(3), 149-163.

Smith, K. S. & R. D. Simpson. (1993). Becoming successful as an international teaching assistant. *The review of higher education, 16*(4), 483-497.

Stapleton, P. (2006) Critiquing research methodology: Comments on broader concerns about complex statistical studies: A response to Ross. *Applied*

Linguistics, 27(1), 130-134.

"STIA: Summer Teaching Institute for Associates" at University of California at Santa Barbara. Retrieved from http://oic.id.ucsb.edu/stia-summer-teaching-institute-associates

Taras, M. (2008). Summative and formative assessment: Perceptions and realities. *Active learning in higher education, 9*(2), 172-192*.*

Taras, M. (2005). Assessment-Summative and formative-Some theoretical reflections. *British journal of educational studies, 52*(4), 466-478.

"TCE Comparisons." (Online policy website.) Retrieved from http://aer.arizona.edu/teaching/reports/compsummary.pdf.

"TCE Guide" (Online policy website). Retrieved from http://aer.arizona.edu/teaching/Guide/TCEGuide.asp.

"Teaching assistant evaluation and improvement handbook." (1997). (Online handbook from University of Wisconsin-Madison, College of Engineering.) Retrieved from http://www.engr.wisc.edu/services/elc/tahand.pdf.

Thomas, C. F., & P. K. Monoson,. (1993). Oral English language proficiency of ITAs: Policy, implementation, and contributing factors. *Innovative higher education, (17)*3, 195-209.

"TOEFL Speaking" (Sample TOEFL tasks from Canadian College of Educators online) Retrieved from http://www.collegeofeducators.ca/onlinepdf/TOEFL%20.pdf

"TOEFL iBT Scores" retrieved from ETS website: www.ets.org/toefl50.html

"TOEFL iBT Tips" retrieved from ETS website: www.ets.org/toefl/tips

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*(4), 713-729.

Tyler, A. (1995). The co-construction of cross cultural miscommunication: Conflicts in perception, negotiation, and enactment of participant role and status. *Studies in second language acquisition, 17*, 129-152.

UA Fact Book 2000-01. (Online OIRPS Website.) Retrieved from http://oirps.arizona.edu/files/Fact_Book/NC_Factbook00_01.pdf.

UA Fact Book 2008-09. (Online OIRPS Website.) Retrieved from http://oirps.arizona.edu/UAFactBook.asp .

"Understanding TCE Results" at the University of Arizona. (Online OIRPS Website) Retrieved from http://oirps.arizona.edu/TCEUnderstandingResults.asp.

Underhill, N. (1987). Testing spoken language: A handbook of oral testing techniques. New York: Cambridge University Press.

Vaden-Goad, R. E. (2009). Leveraging summative assessment for formative purposes. *College Teaching, 27*(3), 153-156.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489-508.

Weimer, M. (2001). Learning more from the wisdom of practice. *New directions for teaching and learning, 86*, 45-56.

Weir, C.J. (1993). *Understanding and developing language tests.* Hemel Hemstead: Prentice Hall.

Wesche, B. (1987). Second language performance testing: Ontario test of ESL as an example. *Language Testing, 4*, 28-47.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*(4), 693-697.

Wilson, R. (2010, Feb 7). For-profit colleges change higher education's landscape. *The Chronicle.* Retrieved from http://chronicle.com/article/For-Profit Colleges-Change/64012/.

Wylie, E. C. and Tannenbaum, R. J. (2006). TOEFL academic speaking test: setting a cut score for international teaching assistants. (ETS RM-06-01). Princeton, NJ: ETS.

Xi, Xiaoming. (2008). Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs. (ETS RR-08-02, TOEFLiBT-03). Princeton, NY: ETS.

Yule, F. & P. Hoffman. (1990). Predicting success for international teaching assistants in a U.S. university. *TESOL Quarterly, 24*(2), 227-243.