Research Memorandum

TOEFL® Academic Speaking Test: Setting a Cut Score for International Teaching Assistants

E. Caroline Wylie

Richard J. Tannenbaum

Research & Development

January 2006
RM-06-01

TOEFL® Academic Speaking Test: Setting a Cut Score for International Teaching Assistants

Caroline Wylie and Rick Tannenbaum

ETS, Princeton, NJ

January 2006

Abstract

The purpose of this study was to establish cut scores for international teaching assistants (ITAs) on the new TOEFL® Academic Speaking Test (TAST), which is the stand-alone equivalent of the speaking section of the TOEFL Internet-based test (TOEFL iBT). Two separate cut scores were established: first, a cut score for minimally acceptable speaking skills in order to have the lowest level of ITA contact with undergraduate students; and second, a cut score to establish a TAST score that corresponds to the Test of Spoken English™ (TSE®) score of 50.

A panel consisting of 18 experts was convened to participate in the standard-setting study. In conducting this study, the panel employed the benchmark method (Faggen, 1994), which is similar to the examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000). As a result of two rounds of judgments with discussion in between, the cut score for the TAST was set as 23 out of 30 scaled score points, and the TSE-50 equivalent score was established as 26 out of 30 scaled score points.

Key words: Authentic language tests, cut scores, international teaching assistants, speaking skills assessment, standard setting, Test of Spoken English (TSE), TOEFL, TOEFL Academic Speaking Test (TAST), TSE-50 equivalent score

Table of Contents

List of Tables

Introduction

The new Test of English as a Foreign Language™ (TOEFL®), known as the TOEFL iBT, became available for students to take in September 2005. The TOEFL iBT is the product of a decade of research at ETS and has some significant changes from the previous version of the test, most notably the inclusion of a speaking section. This new speaking section will also be available as a stand-alone assessment known as the TOEFL Academic Speaking Test, or TAST. (This study focuses only on the speaking section of the TOEFL iBT, and it will be referred to as TAST throughout this report.)

In the past, some institutions have used the Test of Spoken English™ (TSE®) as part of their screening process for international teaching assistants (ITAs). Given that the TOEFL iBT has a speaking component (and that the TAST is available on its own), this standard-setting study was conducted in order to establish a cut score for the TAST in the context of use for awarding international teaching assistantships.

Given the current widespread use of the TSE, it was also desired to understand the potential relationship between scores on the two tests. The second part of the study was to establish a score connection between the TAST and the TSE.

Standard Setting

The process followed to establish cut scores is known as *standard setting*. Standard setting is a general label for a number of approaches used to identify test scores that support decisions about test takers' (candidates') level of knowledge, skill, proficiency, mastery, or readiness. For example, typically, in order for an international student to gain admittance into a North American university where the language of instruction is English, he or she must achieve a certain score (standard) on the TOEFL. This score (or scores, if multiple section-level cuts are used), set by each institution, reflects the minimum level of English language competence the particular institution believes is necessary for a prospective student to function successfully at the institution. The score reflects a standard of readiness to learn subject matter taught in English at that institution. Students with TOEFL test scores at or above the threshold score have demonstrated a sufficient level of English proficiency to study at the institution; those with test scores below the threshold have not yet demonstrated a sufficient level of English language proficiency to study at the institution. A process similar to the one described in this report was used to map TOEFL scores on to the Common European Framework (Tannenbaum & Wylie, 2004).

It is important to recognize that a cut score, or a threshold test score, is typically the outcome of informed expert judgment. There is no absolute, unequivocal cut score. There is no single correct or true score. A cut score reflects the values, beliefs, and expectations of those experts who participate in its definition and adoption, and different experts may hold different sets of values, beliefs, and expectations. Its determination may be informed by empirical information or data, but ultimately, a threshold score is a judgment-based decision.

As noted by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the rationale and procedures for a standard-setting study should be clearly documented. This includes the method implemented, the selection and qualifications of the panelists, and the training provided. With respect to training, panelists should understand the purpose and goal of the standard-setting process (e.g., what decision or classification is being made on the basis of the test score), be familiar with the test, have a clear understanding of the judgments they are being asked to make, and have an opportunity to practice making those judgments. The standard-setting procedure in this study was designed to comply with these guidelines; the methods and results of the study are described below.

The TOEFL Academic Speaking Test (TAST)

The TOEFL Internet-based test (TOEFL iBT) is the result of research conducted by ETS and the TOEFL program to produce a new generation of English language tests and instructional tools.[1] This new generation of assessments includes authentic language and tasks; measures all four language skills (listening, reading, speaking, and writing); includes tasks in which the learner integrates more than one skill; and provides students, teachers, and institutions with more information about the learner’s ability and how he or she can improve.

The TAST (the speaking stand-alone section of TOEFL iBT) consists of six speaking tasks, each of which is scored on a 0-4 scale. The topics vary across the tasks, as does the format. Two tasks require students to speak about familiar, everyday topics; two involve campus situations (such as discussing the impact of a fee increase); and two involve academic course content (i.e., listening to an excerpt from a seminar or lecture and then responding verbally to questions about the content). In terms of presentation format, two tasks require students to respond to a brief written prompt, two tasks require students to first listen to spoken prompt and then respond to it, and two require them to integrate information provided in both written and spoken formats. TAST total scores are reported on a scale that ranges from 0 to 30.

This report is presented in three major sections. The first section describes the panelists who were involved. The second section describes the standard-setting method that was implemented to establish the cut score for ITAs on the TAST, and presents the results. The third section presents the approach used to connect the TAST to the TSE and presents the outcome.

Panelists

The panel was composed of 18 experts from universities across the country. ETS staff familiar with institutions that used the TSE compiled a list that was geographically diverse and that represented both large and small, public and private institutions. Contact was made with each institution, and the project and the type of expertise that a panelist would need to have were described. Each potential panelist submitted a brief biographical form in order to verify that they met the requirements. The panelists were selected for their experience with ITA admissions within the university, their work with testing and/or placing international students in teaching assistant positions, and their familiarity with the existing TSE.

Table 1 presents the demographic characteristics of the panelists, along with information about their institutions. Appendix A provides the panelists' affiliations.

Table 1

Panel Demographics

	Number
Gender	
Female	13
Male	5
Panelist selection criteria[a]	
Faculty advisor of international students	3
Instructor of course(s) in which international students are enrolled	11
Teaching assistant of course(s) in which international students are enrolled	1
Faculty involved in admission decision making	3

(Table continues)

Table 1 (continued)

	Number
Involved in testing/placing/training ITAs	12
Geographical location	
Central	6
West	4
Northeast	5
Southeast	3
Institution type	
State/public	12
Private	6

[a] Some members met more than one criterion, so percentages are not reported.

Establishing a Cut Score on the TAST

Activity Prior to Standard Setting

Before the standard-setting meeting, each panelist was given a homework assignment (see Appendix B) that consisted of two parts. The first part asked panelists to provide a description of the screening process their institution used to identify and support international teaching assistants. These descriptions revealed that a wide range of selection procedures used were being used by the institutions, and that ITAs could be awarded a range of positions that required varying levels of contact with undergraduate students, depending on how the ITA candidate performed in the selection procedure. Thus, based on their responses, for the purpose of this standard-setting exercise, the cut score was discussed in terms of the score that a teaching assistant would need to obtain in order to have the lowest level of speaking contact with undergraduate students and yet still be considered a TA. (See Appendix C for five examples of admissions processes as described by the panelists.)

The second part of the homework assignment asked the panelists to think about critical tasks and skills for speaking, and to write down key indicators that distinguished someone with weak skills from someone with strong skills. They were asked to bring these responses to the study, since they would be helpful for some of the group discussions.

Panelist Training

Panelists were provided with an overview of the purpose of the study and a definition of a cut score as applied to the current purpose. Appendix D provides the agenda that was followed. The cut score was defined as the level of performance on the TAST that reflected the English language ability of a candidate who had a level of English-speaking proficiency adequate for the job of a teaching assistant. In addition, the panelists were provided with an overview of the TAST.

The first major event of the training process had panelists summarizing the key aspects of minimally acceptable English-speaking ability for an international teaching assistant. To facilitate these summarization exercise, panelists were encouraged to refer to their homework notes. This task was completed in small groups. A member of each group recorded on chart paper various aspects of speaking that helped the panelists distinguish weak speakers from strong speakers. The groups then fleshed out the particular identifiers of minimally acceptable speaking. Each group’s charted summary was posted and discussed by the whole group so that the panel had an opportunity to comment and, as appropriate, suggest modifications. This exercise was designed to bring the groups to an agreed upon, shared understanding of the construct of minimally acceptable speaking for a first-year graduate teaching assistant (that is, focusing on the skills needed to be accepted as a graduate teaching assistant, rather than on the skills that one might have after having been in the role for several years). The whole-panel agreed-upon summaries remained posted to guide the standard-setting judgment process. See Appendix E for an example of one group’s chart.

Standard-Setting Process

The standard-setting process applied to the TAST is known as the benchmark method (Faggen, 1994), and is similar to the examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000). As applied to the TAST, the process included the panelists first reviewing the six items of the TAST and the scoring rubric. Operationally, the panelists were asked to read a TAST item and to listen to sample spoken responses to the item that served to illustrate each score point on the rubric (1, 2, 3, 4). No responses were provided to illustrate the 0 score, since that score is reserved for when no response has been attempted or the response is off-topic. The panelists listened to one response per score point. They were asked to consider the difficulty of the English language skill addressed by the item, the language features valued by the rubric, and the skill set of a candidate who would be allowed to work as an ITA. Panelists, independently, were asked to pick the lowest scoring sample response that, in their expert judgment, most appropriately reflected the response of an ITA candidate who had just minimally acceptable speaking ability. This basic process was followed for each of the six TAST items.

Panelists independently completed their judgment for the first TAST item. They were asked to stop, and were given an opportunity to ask questions if they were unsure about the standard-setting purpose or process. No one asked for any clarification. At this point, panelists were formally asked to acknowledge if they understood what they were being asked to do and the overall judgment process. They did this by signing a training evaluation form confirming their understanding and readiness to proceed (an example is provided in Appendix F). In the event that a panelist was not yet prepared to proceed, he [sic] would have been given additional training by one of the ETS facilitators. All panelists signed off on their understanding and readiness to proceed. Panelists independently completed their judgments on the remaining items.

The ETS facilitators computed each panelist's standard-setting judgment for the TAST, summing the scores across the six items for each panelist. The mean cut score across all panelists was computed, as was the median, standard deviation, minimum cut score, and maximum cut score. The cross-panelist summary information was posted (mean, median, minimum, and maximum scores were presented as integer values) and used to facilitate a discussion. Each panelist also had his or her own cut score. The panelists with the minimum cut score and maximum cut score were asked, if they felt comfortable identifying themselves, to begin the discussion. Both panelists readily identified themselves and discussed their judgment processes. The other panelists were then encouraged to share their cut scores and decision rationales. At the conclusion of the group discussion, the panelists were given an opportunity to change their overall cut score if they wanted some aspect of this discussion to be reflected in their final judgment. Having considered each item separately for the first-round judgment and, in so doing, becoming familiar with the demands of the test, the panelists were then asked to consider overall performance for their second-round judgments. The discussion began with a presentation of the mean raw total score, and panelists discussed their decision rationales in relation to the total score. Thus, making their second-round judgments at the overall level was in keeping with nature of the discussion, and panelists were easily able to make the transition. The panelists were reminded that they could keep their first-round cut scores; they were not obligated or expected to change their cut scores. They then recorded their second-round (final) judgments (selecting an integer score value). (See the Appendix G for a copy of the judgment recording form—for first-round and second-round decisions—completed by each panelist.)

Standard-Setting Results

The first-round and second-round judgments for the TAST are presented in Table 1 in Appendix H. Each panelist's individual cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, and maximum). Table 2 presents the summary results for the Round 1 and Round 2 judgments made by the panelists.

Table 2

First- and Second-Round Judgments for the TAST Cut Score

TAST	Round 1 (Raw score)	Round 2 (Raw score)	Scaled score
Mean	18	18	23
Median	18	18	23
Standard deviation	2.11	1.11	
Minimum	15	16	
Maximum	23	21	

The cut score means (and medians) did not change from Round 1 to Round 2 as can be seen in Table 2. The variability (standard deviation) of the panelists' judgments decreased from Round 1 to Round 2, indicating a greater degree of panelist consensus. The second-round mean scores may be accepted as the panel-recommended cut scores for the TAST, once they are transformed to the scaled scores, using a conversion table.

Connecting the TAST and the TSE

Institutions currently using the TSE for awarding ITA positions wanted to understand what TAST scores might mean in relation to the TSE. The second part of this study was designed to address that need. From the responses to the homework assignment that asked the panelists to describe the process their institutions used for selecting ITAs, it was clear that expectations for scores on the TSE varied from 45 to 55 points (on a scale that runs from 20 to 60 in increments of 5 points).

For the purpose of this study, it was decided that only the score of 50 on the TSE would be benchmarked against the TAST. This decision was made by ETS staff members who were familiar with the typical expectations universities set for the TSE, and the decision was further borne out by the panelists' review, which indicated that required TSE scores ranged from 45 to 55. The task presented to the panelists was to determine what the likely TAST score would be for a hypothetical candidate who received a 50 on the TSE. In order to complete this task it was critical that the panelists were familiar with the TSE items, the scoring rubric, and panelists' performances at the critical score points.

The panelists were first given an opportunity to review the nine TSE items in conjunction with the scoring rubric. For each item, they then listened to three candidates' responses: one that scored a 40, one that scores a 50, and one that scored a 60 for that particular item. The panelists then discussed the characteristics of a candidate who would score a 50 across the nine items, and these were noted on chart paper (see Appendix I). Up to this point, the conversation focused exclusively on the TSE, understanding its demands, and what a score of 50 would mean in terms of what a candidate could and could not do.

The candidates were then shown the final judgment form and the question, "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?" The panelists were then referred to the charts that listed the features of a speaker with minimally acceptable skills, and they were reminded that those descriptions resulted in a TAST cut score of 18 points. One approach to thinking about the second standard-setting question was to consider whether the descriptors of what a candidate with a TSE score of 50 could do in terms of speaking seemed similar to, more skilled than, or less skilled than the descriptors of an ITA with minimally acceptable speaking skills. There was some discussion about the difference between the tests, one of which was that the TSE assessed speaking skills across a wider range of contexts than the TAST although the integrated listening-speaking and reading-listening-speaking TAST items seemed more demanding.

The panelists were then asked if they understood the purpose or process for this second standard setting. The group asked some procedural questions but quickly indicated that they understood the task. As soon as everyone verbally indicated their understanding they were directed to make their individual judgments.

The ETS facilitators computed the mean cut score across all panelists, as well as the median, standard deviation, minimum cut score, and maximum cut score. In addition, a count of how many selected 18 points (the minimally acceptable TAST cut score), less than 18 points, and greater than 18 was provided for the group.

The cross-panelist summary information was posted and used to facilitate a discussion. The panelists with the minimum cut score and maximum cut score were asked to begin the discussion, with other panelists encouraged to share their cut scores and decision rationales. At the conclusion of the group discussion, the panelists were given an opportunity to change their overall cut score if they felt that they wished to reflect some aspect of the discussion in their final judgment. Panelists were reminded that they could keep their first-round cut scores; they were not obligated or expected to change their cut scores. Panelists then recorded their second-round (final) judgments.

Standard-Setting Results

The first-round and second-round section-level judgments are presented in Table H2 in Appendix H. Each panelist's individual cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, and maximum). Table 3 presents the summary results for the panelists' Round 1 and Round 2 judgments.

Table 3

First- and Second-Round Judgments for the TAST Score That Relates to the TSE 50

TAST	Round 1 (Raw score)	Round 2 (Raw score)	Scaled score
Mean	20	20	26
Median	19	20	26
Standard deviation	1.79	1.20	
Minimum	16	18	
Maximum	24	23	

Similar to the previous series of judgments, the cut score means did not change from Round 1 to Round 2, as can be seen in Table 3. However, the variability (standard deviation) of the panelists' judgments decreased, indicating a greater degree of panelist consensus. The second-round mean scores may be accepted as the panel-recommended TSE-50 equivalent scores for the TAST, once they are transformed to the scaled scores.

Conclusions

The purpose of this study was twofold:

1. To determine a cut score for ITA selection on the TAST
2. To establish a TAST score that corresponded to the TSE score of 50

A panel of 18 experts participated in the standard-setting study, and used the benchmark method (Faggen, 1994)—also referred to as the examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000)—to answer the first question. As a result of their two rounds of judgments and the between-round discussion, the cut score for the TAST was set as 23 out of 30 scaled score points.

One common approach in standard-setting studies is the inclusion of more than one round of item-level judgments, with discussions between rounds (Busch & Jaeger, 1990). The rationale for such discussion—which may or may not be accompanied by normative data, such as proportion correct values—is that panelists have the opportunity to hear and consider other relevant perspectives, which they can then incorporate into their next round of item-level judgments. The inclusion of discussion tends to result in higher cut scores and reduced variability (Hurtz & Auerbach, 2003).

Both parts of the study included two rounds of judgments, with between-round discussion. For the first question, setting the TAST cut score, the second round of judgments used an approach that had been employed in a previous standard-setting study (Tannenbaum & Wylie, 2004): The focus of the judgments between rounds shifted from the item level (Round 1) to the domain or construct level (Round 2). The first-round item-level judgments were important and necessary for engaging the panelists in considerations of language demands posed by each of the six TAST items. But given the holistic nature of speaking skills, it was believed to be more meaningful and appropriate for the second round of judgments to be framed in terms of overall performance. Once panelists understood the item content and received feedback about their initial TAST cut score and the panels' cut scores (computed from the item judgments), the stage was be set for meaningful discussion at the domain or construct level; hence, it was believed more meaningful and relevant to make postdiscussion judgments at that same level, rather than deconstructing the domain, in essence, by repeating item-level judgments during the second round.

For the second question, establishing a correspondence between the TAST and the TSE, a decision was made to use a judgmental standard-setting approach rather than employing an empirical approach of having a large number of students take both tests in order to derive equivalent scores. This decision was driven in part by the need to determine comparable score values in a relatively compressed time frame, making large-sample recruitment logistically infeasible. It was also desired to involve stakeholders and experts in the process of aligning scores on TSE and TAST, rather than doing so in an empirically driven approach. The standard-setting approach afforded stakeholders hands-on experience with the new TAST and engaged them in open discussion of the new test. The panel considered responses to the TSE items that were scored as 40, 50, and 60 points; created a description of a hypothetical candidate who would score 50 points on the test; and then made a professional judgment regarding the likely TAST score that this hypothetical candidate would receive. As a result of their two rounds of judgments and the between-round discussion, the TSE-50 score on the TAST was established as 26 out of 30 scaled score points.

One group member noted during the discussion period that the initial cut score on the TAST was established in the context of minimally acceptable or just satisfactory; whereas the score of 50 on the TSE was seen as a safe score and thus represented a performance slightly above what had been defined previously as minimally acceptable. While not every panelist necessarily saw the TSE-50 in exactly the same way, following this observation there was strong consensus among the other panelists that there was coherence between the two separates cut scores that they had established for the TAST: It was reasonable to expect that the TSE-50 cut score on the TAST should be higher than the just-previously established minimally acceptable TAST cut score.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, *27*, 145-163.

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

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*, 355-366.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*, 584–601.

Hurtz, G. M., & Hertz, N.R. (1999). How many raters should be used for establishing cutoff scores with the Angoff Method? A generalizability theory study. *Educational and Psychological Measurement*, *59*, 885-897.

Tannenbaum, R. J., & Wylie, E. C. (2004). *Mapping test scores onto the Common European Framework* (TOEFL Research Rep. No. RR-80, ETS RR-05-18). Princeton, NJ: ETS.

Notes

[1] A full description of the TOEFL 2000 research framework is available in the TOEFL monograph series, numbers 15–20. The monographs can be downloaded for free in PDF format from the research section of the TOEFL Web site at www.ets.org/toefl/research.

List of Appendixes

Appendix A

Panelists' Affiliations

Standard-Setting Participants

Name	Affiliation
Richard Burnson	University of Wisconsin-Madison
Vicki Bergman-Lanier	Spring International Langauge Center, University of Arkansas
Julia Cayuso	University of Miami
Linda DiPietro	Indiana University
Tammy Guy Harshbarger	English Language Programs, University of Washington
Gene B. Halleck	Oklahoma State University
Jane Kenefick	American Language Program, Columbia University
Joseph W. Matterer	English Language Institute, University of Delaware
Patricia Pashby	University of Oregon
Barbara Schroeder	Princeton University
Doris Yaffe Shiffman	Johns Hopkins University
Marilyn Seid-Rabinow	University of California at Berkley
Martha Stacklin	University of California, San Diego
Christos Theodoropulos	Drexel University
Julie E. Vance	Yale University
Elizabeth Wittner	University of Virginia
Lawrence J. Zwier	Michigan State University

Note. Permission was asked of all panelists to publish their names and affiliations. One panelist did not wish to be listed in the final report.

Appendix B

Homework Task

Standard Setting on the TOEFL Academic Speaking Test (TAST)

We will soon be meeting to discuss the level of speaking skills you believe are necessary for international students to demonstrate in order to work as teaching assistants. We will accomplish this by reviewing the TOEFL Academic Speaking Test (TAST) that your university may use to measure this skill set in international graduate students applying for *teaching assistants' positions.*

During the meeting we will decide the minimum scores on this test that you, and your colleagues participating in this study, believe reflect the levels of English language speaking ability necessary for an entering international student to deal satisfactorily with demands and expectations of your university for this role. As part of the study process, we will ask you to share your experiences with an international student or students with whom you have interacted recently.

In preparation for the meeting we ask that you please complete two brief tasks:

1. Briefly describe the selection process and nature of support for international teaching assistants at your institution. Please e-mail your response to Caroline Wylie, ecwylie@ets.org, by September 15, 2004.
2. Please complete the attached exercise. It has been designed to get you thinking about the kinds of speaking tasks graduate/professional students at your university are expected to complete in the role of a teaching assistant and the speaking skills they need. We ask that you bring your responses to the meeting, as this information will greatly facilitate our discussion.

Think about the range of activities for all graduate/professional teaching assistants at your university that require speaking. Below are two examples of speaking activities in which teaching assistants are likely to be engaged. Please complete the table by adding other *tasks or activities in which oral communication* is important for graduate/professional students.

Speaking Tasks for a Teaching Assistant
a) Talk with professors about lecture material prior to working with students
b) Arrange seminar times with students
c)
d)
e)
f)
g)
h)
i)

Think of a particular *international teaching assistant* with whom you interacted who you thought was a *good English-language speaker*. Think of an international teaching assistant that you thought was a *poor English-language speaker*. Write down the reasons why you think he/she is a good or a poor speaker.

A Good Speaker	A Poor Speaker

Appendix C

Selected Responses to the Homework Task

Response A

Academic units select ITAs. Prospective ITAs must obtain a minimum score of 50 on the TSE or SPEAK tests, or pass a university performance test. New ITAs must also attend the ITA Orientation (1 1/2 days); and the All-Campus TA Orientation (for domestic and international TAs) or a teaching orientation sponsored by their academic unit (generally 2 days, including a videotaped microteaching). The Center for Teaching Excellence and various academic units provide pedagogical support during the semester for TAs and ITAs in the form of seminars, workshops, staff meetings, one-on-one consulting, etc. Students who do not receive a minimum of 50 on the SPEAK or TSE are directed to two ESL courses for ITAs, or they may seek private tutoring at their own expense.

Response B

The requirements for ITAs to be eligible for any TA assignments are as follows: a TOEFL score of 550-600 depending on graduate program, a SPEAK Test score of 55-60 or a minimum score on the university's Performance Test, and/or a satisfactory completion of the ITA Summer Orientation Program with a final recommendation of AA (all assignments). All tests are offered free of charge to graduate students applying for teaching assistantships and are administered through the English Language Center (an intensive English program). In terms of ITA support, we offer language counseling year round; an ITA fall course, Classroom Communication Skills, which focuses on pronunciation and intelligibility; and an intensive ITA summer preparation program. The summer program is a four-week program before the fall term which covers language, language of pedagogy, and the culture of the U.S. classroom. All new ITAs regardless of language proficiency are required to attend. Exemptions are made only to those with sufficient language proficiency who also have attended a university in the U.S as undergraduate students. Average program size is 25-35 students. The program includes free tuition to participants, free housing in the dormitories for the program duration, books, and a living stipend of $500 for full attendance for five weeks.

Response C

ITAs have to pass TSE or SPEAK (with a minimum score of 50) in order to be eligible to take the mock-teaching test (The ITA Test) through which they earn certification for classroom teaching. If a student gets a 45 on SPEAK/TSE she or he can take a remedial, one-credit course that prepares students for these tests. With a score of 40 or below they need to find a tutor, since we have found that our course really is not enough for people at such a low level. If a person does not get a passing score on the mock-teaching test we also have a remedial course to help him or her prepare for the next administration.

Response D

Concerning the selection process for ITAs, the individual grad departments select the candidates based on their qualifications and a minimal TOEFL score of 250. One department which uses a lot of international TAs requires a phone interview during which an abbreviated form of the SPEAK test is used. In order to be a TA, all international students must complete a training program, for which they receive a stipend. This program has been offered for about twenty years. At the end of the program, the students are given the SPEAK test and an institutional instructional assessment. The scores on the two tests are combined and fall into six categories. The categories indicate what types of instructional duties the ITA is allowed to perform. Students who do not score sufficiently to be a TA do not lose their funding the first year. In addition, oral intelligibility classes are provided for these students during the fall semester, and the students can retest at the end of the semester. All of this support and the training program and classes are funded by the Office of the Provost and overseen by the Office of Graduate Studies. My unit, the English Language Institute, provides the instruction in the program and does all of the testing.

Response E

The graduate office of the university processes applications and sends them on to the departments. The ESL Program for Int'l TAs is not involved in the selection process. Each department chooses its own students from the applicants and sets its own standards for the TOEFL. Departments vary in terms of what funding is available to their incoming students and therefore what jobs are assigned to them. Some departments wait until students are in their second year at the university before asking them, both international and native, to be TAs. Others give just the international TAs a year as graders before they are assigned to work with undergraduates face-to-face. (Graders are, nevertheless, considered to be TAs and are funded as such.) Still other departments must put all incoming graduate students in the classroom in some form, e.g., as lab assistants, classroom teachers (for language classes) or conference section leaders, regardless of the level of English proficiency of the ITAs. There are departments too which don’t need TAs and have enough funding for research assistantships for all of their students. On the whole all Ph.D. candidates are funded their first year as TAs or RAs. Once the graduate students are at the university, a member of the department usually decides on the assignments, depending on various factors, including language proficiency judged by simply by talking to each student. No ESL professionals are involved in this process. The instructor for the ESL program for ITAs tests students at the beginning of the year to decide what coursework, if any, a prospective or current new ITA needs to improve his or her communication skills. Only on rare occasions is the instructor consulted before the department decides who will be its TAs from the new pool of international graduate students. On the other hand, departments that can defer teaching till the second year for ITAs do take into consideration the letters from ESL instructor when they make their decisions about teaching for the coming year. The ITA program offers two one-semester courses. The first one emphasizes speaking, listening, and pronunciation for the classroom. The second one, for students who score higher, emphasizes teaching skills, cultural knowledge of the classroom, and still more language practice. Students who demonstrate a high level of English proficiency and good classroom communication skills are excused from the course sequence. As a result of this system, students who might have scored high on the TSE could still be asked to enroll in the second course, which emphasizes teaching skills. On the other hand, students whose oral skills are weak might still end up as TAs while they are taking the first-level ESL class.

Appendix D
Agenda

Standard Setting Study for TAST

September 24, 2004

Doubletree Hotel - Philadelphia, PA

Maestro Meeting Room

Agenda

Morning Schedule	
8:00 – 8:30	**Continental Breakfast**
8:30 – 8:45	Introductions
8:45 – 9:00	Purpose of the study and brief overview of TAST
9:00 – 9:30	More detailed review of the TAST and Rubrics
10:00 – 10:15	**Break**
10:15 – 11:00	Minimally acceptable speaking skills for an ITA
11:00 – 11:30	Overview of standard setting method
11.30 – 12:00	Practice making 1st judgment and discuss
Afternoon Schedule	
12:00 – 1:00	**Lunch – Academy Café (Second Floor)**
1:00 – 1.45	1st round judgments on remaining items
1:45 – 2:00	Break
2:00 – 2:30	Discussion and final round judgments
2:30 – 3:00	Review the TSE and Rubric
3:00 – 3:30	Define the meaning of a TSE score of 50
3:00 – 4:00	Connecting the TSE to the TAST
4:00 – 4:15	**Break**
4.15 – 4.45	Discussion and final round judgments
4:45 – 5:00	Wrap up and adjourn

Appendix E

Example of Panel Summaries of Language Skills

Minimally Acceptable Skills

- Some ability to compensate in cases of misunderstanding
- Hits general semantic territory without taking too long
- Stresses most key words in thought groups
- Segmental errors not distracting
- Basic intonation patterns present
- Minimal word order problems
- Uses word forms of key words accurately enough not to distract
- Discourse markers give an adequate guide
- Displays awareness of particular audience in a particular situation

The group discussed weak skills and strong skills as range finders, but only charted the minimally acceptable skills.

Appendix F

Training Evaluation Form

TAST Standard Setting

ID: ___________________________

Please indicate your level of understanding regarding each of the following tasks.

(A rating of “insufficient” means you still have unanswered questions and are not ready to begin making standard-setting judgments. A rating of “sufficient” means the training and discussion answered your questions and you are ready to begin the standard-setting process.)

TASK	SUFFICIENT	INSUFFICIENT
Develop the concept of the candidate with minimally acceptable speaking ability		
Understand the steps in the standard-setting process for the TAST		

I need additional information about the concept of the candidate with minimally acceptable speaking ability, and/or the process of standard setting before I am ready to begin the Standard-Setting Process.

No □ Yes □ If yes, list specific information needs below:

____

____

____

I now have the information I need to begin the Standard-Setting Process. No □ Yes □

________________ (Date)

________________________________ (Signature)

________________________________ (Print name)

Appendix G

Judgment Form

Round 1 Judgments

Item	Circle the score that a candidate with *minimally acceptable speaking* would achieve on each item.			
1	1	2	3	4
2	1	2	3	4
3	1	2	3	4
4	1	2	3	4
5	1	2	3	4
6	1	2	3	4

Do Not Write in this Space.

	End of Round 1 cut-score
My initial recommended cut-score (range 4 – 24)	
Group average	

Round 2 Judgments

	Write the overall score that a candidate with minimally acceptable speaking skills would achieve on the TAST.
My final recommended cut-score (range 4 – 24)	

____

(Signature)

Part II: Mapping the TSE Score of 50 to the TAST

Group average cut score for the TAST (given the skills that a minimally acceptable speaker would need for ITA role)	Group average = ______ out of 24 points

Round 1

Given the descriptions of what a candidate with a score of 50 on the TSE could do, how would that candidate perform on the TAST?

A candidate with a TSE score 50 would score ______ out of 24 points on the TAST.

Please provide a rationale below for your selection:

Round 2

Given the descriptions of what a candidate with a score of 50 on the TSE could do and the discussion, how would that candidate perform on the TAST?

A candidate with a TSE score 50 would score ______ out of 24 points on the TAST.

Please provide a rationale below for your selection (if your score did not change just note "as above"):

Appendix H

First- and Second-Round Judgments

Table H1

Judgments for the TAST

Raw data	Round 1	Round 2
P1	15	16
P2	18	17
P3	18	17
P4	18	18
P5	17	17
P6	15	18
P7	15	18
P8	23	18
P9	18	18
P10	16	17
P11	19	19
P12	17	17
P13	19	18
P14	17	17
P15	20	21
P16	19	18
P17	21	19
P18	17	17
Summary		
Mean (truncated)	18	18
Median (truncated)	18	18
Standard deviation	2.11	1.11
Minimum	15	16
Maximum	23	21

Table H2

Judgments for the TSE Equivalent

Raw data	Round 1	Round 2
P1	21	20
P2	19	20
P3	16	19
P4	20	20
P5	19	19
P6	24	23
P7	18	22
P8	21	21
P9	17	18
P10	18	20
P11	19	20
P12	19	19
P13	19	19
P14	20	20
P15	20	21
P16	21	21
P17	21	21
P18	19	20
Summary		
Mean (truncated)	20	20
Median (truncated)	19	20
Standard deviation	1.79	1.20
Minimum	16	18
Maximum	24	23

Appendix I

Discussion of the Meaning of the TSE 50

- Reasonable speed, well paced
- Appropriate register
- On-target word choice
- Mostly easy to understand (both organizationally and in terms of pronunciation)
- Range of complex grammar structures
- Minimal hesitations, ability to self-correct, not as confident as a “60”
- Key verbal signals with appropriate pauses
- Grammar errors not distracting
- Content well developed
- Good vocabulary, stress on key words
- Personalized, not “scripted” responses
- Good audience awareness