Validating the Use of TOEFL® iBT Speaking Section Scores for ITA Screening and Setting Standards for ITAs

Xiaoming Xi

Educational Testing Service

The Test of English as a Foreign Language ™ (TOEFL®) has undergone major revisions, including the introduction of speaking as a mandatory section on the TOEFL Internet-based test (iBT). The TOEFL iBT Speaking Section has been designed to measure a candidate's ability to communicate orally in English in an academic environment. Although it is used primarily to inform admission decisions regarding international applicants at English medium universities, it may also be useful as an initial screening measure for international teaching assistants (ITAs).

As reviewed in Plakans & Abraham (1990), three major types of tests have been used to test the oral skills of ITAs: the Test of Spoken English (TSE) or SPEAK developed by Educational Testing Service (ETS), oral interviews, and teaching simulation tests. These tests have served complementary functions in ITA testing. Traditionally, the TSE, administered in test centers around the world, has served the purpose of pre-arrival screening. However, while the TSE uses speaking tasks that are contextualized in more general communicative settings, The TOEFL iBT Speaking Section has been designed specifically to measure oral communication skills for academic purposes. Thus, it may be a more appropriate measure for ITA screenings than the TSE, given its focus on academic contexts. In addition, the TSE has mostly been phased out with the launch of the TOEFL iBT test world wide, and a new pre-arrival screening test is needed.

Locally administered SPEAK exams, which use retired TSE forms, have been widely used as an on-site ITA screener, alone or along with locally developed teaching simulation tests. Although the TOEFL iBT test has been launched in the majority of locations worldwide, the SPEAK test can still be used for on-campus initial screening. It should be noted, however, that ETS no longer supports or carries this product. Nevertheless, for incoming international students who submit their TOEFL iBT scores with their applications (including their TOEFL Speaking Section scores), the TOEFL iBT Speaking Section scores could potentially be used for pre-admission screening. Such an approach would aid in identifying candidates who are ready to

teach as well as help determine who needs to be tested using the local test before and/or after they have arrived.

The goals of this study are to provide criterion-related validity evidence for ITA screening decisions based on the TOEFL iBT Speaking Section scores and to evaluate the adequacy of using the scores for TA assignments. First, this study investigates the relationships between scores on TOEFL iBT Speaking Section and scores on criterion measures, intending to establish some association between them. Then, it illustrates how cut scores for TA assignments can be determined based on students' performances on the TOEFL iBT Speaking Section and on the criterion measures.

In this study, two types of criterion measures for the TOEFL iBT Speaking Section were used: locally developed teaching simulation tests used to select ITAs and ITA course instructors' recommendations of TA assignments. Institutions which adopt fairly established procedures to select ITAs were selected. In particular, they use various performance-based tests that attempt to simulate language use in real instructional settings. This type of teaching simulation test was considered to be more authentic in resembling the real-world language use tasks and in engaging the underlying oral skills required in instructional settings, in comparison to a tape-mediated general speaking proficiency test and oral interview (Hoejke & Linnell, 1994). At these participating institutions, various studies have been conducted to support the validity of their tests for ITA screenings or procedures have been established to check the effectiveness of the ITA test for ITA assignments. Whenever feasible, the reliability of the scores on a local ITA test was estimated in this study and then the observed correlation between the local ITA test scores and the TOEFL iBT Speaking Section scores was corrected for score unreliability to reveal the "true" relationships between them. Otherwise, measurement errors associated with scores on both the TOEFL iBT Speaking Section and the local ITA test may disguise the true relationships between them.

The most important focus of this paper is to illustrate the process of setting cut scores for ITA screenings. This involves both methodological considerations and value judgments. On the methodological side, it demonstrates how the overall effectiveness of TOEFL iBT Speaking Section scores in classifying TA assignments can be established by using binary or ordinal logistic regression (Agresti, 2002; Hosmer & Lemeshow, 2000; Menard, 2001). It also discusses two types of errors that may occur when using TOEFL iBT Speaking Section scores for

classifying students for teaching assignments, taking into account their trade-offs, which reflect value judgments, in order to establish an appropriate standard in ITA screening.

Trade-off of Different Classification Errors in Using TOEFL iBT Speaking Section Scores for TA Assignments

When TOEFL iBT Speaking Section scores are used to classify students for TA assignments, two types of classification errors are likely to occur: false positives and false negatives. In this context, false positives occur when those who are not qualified TAs based on their local ITA test scores are predicted to be qualified by their TOEFL Speaking Section scores. In contrast, false negatives occur when candidates who are qualified TAs are predicted to be unqualified by their TOEFL iBT Speaking Section scores. Since ITA programs are gatekeepers for quality undergraduate education, false positives may have more serious impact, since having unqualified ITAs in classrooms may compromise the quality of undergraduate education and infringe on the interests of undergraduate students who pay high tuitions and fees.

The other factor to consider in setting cut scores is to what extent a specific type of error could be rectified. This study examines the use of TOEFL iBT Speaking Section scores as an initial screening measure to help identify qualified TAs. If an unqualified TA were classified as qualified by his/her TOEFL iBT Speaking Section score (a false positive), there would be no way to rectify this error. However, if an otherwise qualified ITA were predicted as unqualified (a false negative), he/she would still have a chance to be tested using the local ITA test once they arrived. The impact would be that his/her TA employment may be delayed until he/she passes the local test. After considering the potential impact of the two types of errors and how rectifiable they are, it was decided that it is more important to minimize false positives at the expense of false negatives.

The study

Four universities participated in this study: University of California, Los Angeles (UCLA), University of North Carolina, Charlotte (UNCC), Drexel University (Drexel), and University of Florida at Gainesville (UF). At all these universities, an in-house ITA screening test has been used alone or in conjunction with the SPEAK test to screen ITAs. At each institution, students who signed up for their local ITA tests were invited to take the TOEFL iBT Speaking Section as well. Table 1 summarizes the data collected at the four institutions.

Table 1 Data Collected at Each Participating school

	TOEFL iBT Speaking Section	In-house ITA test	SPEAK	Instructor recommendations
UCLA	X	X		
UNCC	X	X		
Drexel	X	X	X	X
UF	X	X	X	

The next section uses UCLA as an example to demonstrate the relationship between the local ITA test scores and TOEFL iBT Speaking Section scores and the process of setting cut scores on TOEFL iBT Speaking Section for ITA selection.

Illustrative example -- UCLA

Local ITA Assessments and Requirements for ITAs

The Test of Oral Proficiency (TOP) has recently replaced SPEAK at UCLA for screening ITAs. It is a locally developed test that consists of three tasks: A self-introduction (not scored), a short-presentation on some typical classroom materials provided, and a prepared presentation about a basic topic in the examinee's own field.

The short presentation and the prepared presentation tasks are each double scored in an analytic fashion on Pronunciation, Vocabulary/Grammar, Rhetorical Organization, and Question Handling. The composite TOP score is derived by summing the four scores, with a 1.5 weight assigned to pronunciation. Then it is scaled to a range of 0 to10. A score of 7.1 or higher is necessary for a "clear pass" which will allow a student to work as a TA. A score of 6.4 to 7.0 is considered a "provisional pass", and students receiving scores in this range are required to take an ITA oral communications course prior to or during their first quarter of TA work. A score lower than 6.4 is not high enough to qualify for TA work.

Participants and Procedure

Eighty-four international graduate students who were roughly representative of the TOP examinee population at UCLA took both the TOP and TOEFL iBT Speaking Section. Forty-two (50.0%) of them were classified as clear passes, 15 (17.9%) as provisional passes and 27 (32.1%) as non-passes based on their TOP scores.

Relationships between TOEFL iBT Speaking Section scores and TOP scores

Table 2 demonstrates that the correlations among TOEFL iBT Speaking Section scores and TOP composite and analytic scores were moderately high. After correcting for score unreliability, the correlation between the TOEFL iBT Speaking Section and TOP composite scores was .84. The disattenuated correlations, which were observed correlations corrected for score unreliability, also show that the TOEFL iBT Speaking Section scores had strong correlations with the TOP analytic scores, showing the strongest relationship with the TOP Grammar & Vocabulary scores (.86).

Overall effectiveness of using TOEFL iBT Speaking Section scores for ITA screening

Sixty-five cases, randomly selected from the whole sample, were used in model building and the remaining 19 cases were used in testing the classification accuracy. An ordinal regression model with a logit link satisfied the assumption of parallel

Table 2 Observed and Disattenuated Correlations between the TOEFL iBT Speaking Section Scores and TOP Scores (N=84)

	TOP	TOP Pronunciation	TOP Vocabulary & Grammar	TOP Organization	TOP Question Handling
TOEFL iBT Speaking Section	.78 **.84**	.75 **.81**	.75 **.86**	.68 **.80**	.69 **.82**

Note: The disattenuated correlations are in bold face.

regression lines and also provided good classification results. The results show that the TOEFL iBT Speaking Section scores were a significant predictor of the TA assignment outcomes.

The classification accuracy further demonstrates how the TOEFL iBT Speaking Section scores performed in classifying students into one of the three outcomes. In Table 3, cases on the diagonal were correctly classified and the off-diagonal ones represent incorrectly predicted cases. The model did a superb job of correctly classifying the clear passes (97.0%), fairly well with the non-passes (81.8%), but not as well with the provisional passes (30.0%). This may be due to the fact that the model employed many fewer cases in the provisional pass category. Further, these provisional pass students were borderline students and may be more difficult to classify accurately.

Table 3 True versus Predicted Outcome Categories at UCLA

True TA assignment outcome	Predicted TA assignment outcome			Percentage correct
	Non-pass	Provisional pass	Clear passes	
Non-passes	18	1	3	81.8%
Provisional passes	5	3	2	30.0%
Clear passes	1	0	32	97.0%
Total			Overall percentage	81.5%

Setting the Cut Scores

In the ROC curve for provisional passes (Figure 1), the area under the curve was very high (.91), indicating that the probability of the TOEFL iBT Speaking Section score of a marginal or clear pass student exceeding that of a non-pass student was 91%. That is to say, if we randomly select a clear pass student and a non-pass student, 91% of the time, the TOEFL iBT Speaking Section score of the former will be higher than that of the latter. Table 4 contrasts the true positive and false positive rates for different TOEFL iBT Speaking Section score points for provisional passes. When the cut score is set at 24, no false positives will occur, but the true positive rate will stand at 53.5%. In other words, the model has to misclassify 46.5% of the marginal or clear passes as non-passes to correctly classify all non-passes. If 23 is chosen as the cut score, approximately five out of 100 non-passes may be classified as provisional passes. However, 11.6% (65.1% - 53.5%) more provisional passes will be correctly classified. This would reduce the number of students to be tested locally using the TOP but increase the number of students in ITA training classes. The slightly lower cut score (23) might be justified for two reasons: 1) Many science departments who hire the most ITAs are in dire need of TAs and a larger pool of eligible ITAs would help meet this need; 2) ITA course instructors can offer extra help in class to rectify the situation where non-passes are assigned TA work with concurrent English coursework on oral communication skills.

Using a table similar to Table 4, but for clear passes, 27 was estimated as the optimal cut score for identifying clear passes.

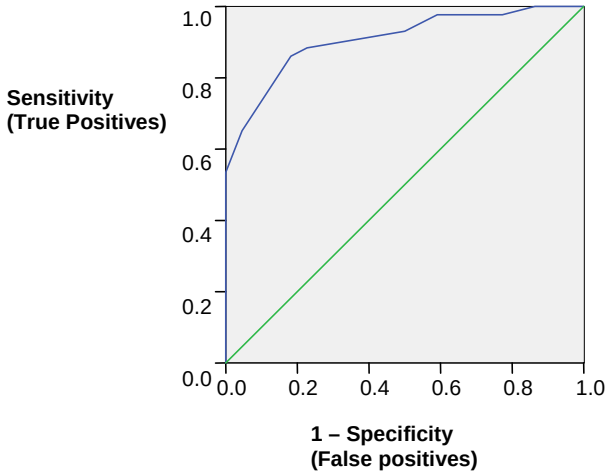

***Figure 1*. The ROC Curve for Provisional Passes with TOEFL iBT Speaking Section Scores as the Predictor**

Table 4 True Positive Versus False Positive Rates at Different TOEFL iBT Speaking Section Cut Points for Provisional Passes

Positive if Greater Than or Equal To	True positive	False positive
18.50	.884	.227
19.50	.860	.182
21.00	.791	.136
22.50	**.651**	**.045**
23.50	**.535**	**.000**
25.00	.395	.000
26.50	.279	.000

Note1: Not all possible cut points are displayed.

Note 2*:* The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All other cutoff values are the averages of two consecutive ordered observed test values. An integer cutoff value such as 21 is possible when the two consecutive test scores in the sample are 20 and 22. The cutoff values are rounded off to integers in the discussion of cut scores in the text because integer scaled scores are reported for the TOEFL iBT Speaking Section.

Cross-validation of the Classification Accuracy

The cut scores derived from the training sample were validated using the independent sample. As shown in Tables 5 and 6 using 23 or 24 as the cut score for provisional passes and 27 for clear passes, the classification accuracy with the independent sample was fairly similar: all

the non-passes were correctly predicted; only one of the provisional passes was incorrectly classified as a clear pass. However, some students were incorrectly classified into the lower categories. This is acceptable given that the false non-passes could be tested again using the local test once they arrive and those false provisional passes can get out of the ITA coursework at the recommendation of the instructors.

The false positive case causes some concern; however, UCLA allows provisional pass students to teach with concurrent enrollment in an English oral communication class. Therefore, if a mechanism is established for ITAs to receive some language support if it is found necessary after they start to teach, it should be reasonable to keep the cut score of 23 for provisional passes.

Table 5 Classification Rate on an Independent Sample with 27 on the TOEFL iBT Speaking Section for Clear Passes and 24 for Provisional Passes

True TA assignment outcome	Predicted TA assignment outcome			Percentage correct
	Non-passes	Provisional passes	Clear passes	
Non-passes	5	0	0	100.0%
Provisional passes	4	0	1	0.0%
Clear passes	4	3	2	22.2%
Total			Overall percentage	36.8%

Table 6 Classification Rate on an Independent Sample with 27 on the TOEFL iBT Speaking Section for Clear Passes and 23 for Provisional Passes

True TA assignment outcome	Predicted TA assignment outcome			Percentage Correct
	Non-passes	Provisional passes	Clear passes	
Non-passes	5	0	0	100.0%
Provisional passes	4	0	1	0.0%
Clear passes	3	4	2	22.2%
Total			Overall percentage	36.8%

Summary of the results from four institutions

Association between TOEFL iBT Speaking Section and local ITA test scores

This study investigated the criterion-related validity of the TOEFL iBT Speaking Section scores for screening ITAs by examining its relationships with the local ITA test scores. The

findings support the use of the TOEFL iBT Speaking Section scores for ITA screening because the TOEFL iBT Speaking Section scores were reasonably correlated with scores on the local ITA screening measures.

As shown in Table 7, the TOEFL iBT Speaking Section scores had the strongest relationship with the UCLA TOP test scores and the UNCC Non-content Based Presentation test, less strong relationships with the Drexel DIP test scores and the UNCC Content-based Presentation test, and the weakest relationship with the UF Teach Evaluation scores. However, due to unavailability of data in some cases, some disattenuated correlations could not be estimated (e.g., Drexel). In other cases, the disattenuated correlations were underestimated as a result of the reliability of the local ITA tests being overestimated. Due to the particular assessment designs, such as single ratings of tasks or using a single task in an assessment, it was not possible to obtain appropriate reliability estimates that would take account of all potential sources of error (e.g., UNCC, Drexel and UF). In yet other cases, the restricted range of scores rendered the observed correlations lower than they would be if the whole range of possible scores were used (e.g., UF). Therefore, the disattenuated correlations provided only a partial picture of the "true" strengths of the relationships among these measures.

Table 7 Correlations between the TOEFL iBT Speaking Section and the Local ITA Test Scores

	UCLA TOP test	UNCC non-content presentation test	Drexel DIP test	UNCC content-based presentation test	UF Teach Evaluation
TOEFL iBT Speaking Section	.78 **.84**[1]	.78 **.93**	.70 Not available	.53[2] **.58**	.44 **.72**

Note 1: The disattenuated correlations are in bold face.
Note 2: This correlation was based on a very small sample (N = 23) and should be interpreted with caution.

The strengths of the relationships were certainly affected by the extent to which the local ITA tests engaged and evaluated non-language abilities. As is evident in Table 8, the criterion measures used in this study certainly represent a continuum of less to more authentic tests. SPEAK can be placed on the left end of the continuum, since it uses tasks that are the least authentic in eliciting speech characteristic of language use in academic settings. The UCLA TOP test, the UNCC Presentation tests and the Drexel DIP test represent fairly authentic

performance-based assessments that simulate the communication typical TA duties involve. On the right end of the continuum is the UF Teach Evaluation, which is an evaluation of videotaped ITAs' actual classroom teaching sessions. The further to the right, the more entangled speaking abilities are with teaching skills, increasing the chances that examinees' speaking abilities are impacted by their teaching skills, and making it difficult for the assessors to separate them out in their evaluations. The scoring rubrics of these local tests also range from primarily linguistically driven criteria to real-world criteria. For example, the scoring rubric for the UCLA TOP test is most representative of a linguistically driven rubric in which teaching abilities are clearly not scored whereas the rubrics for the other three local ITA tests contain, to varying degrees, teaching abilities or demonstration of an understanding of the American university classroom culture. In the latter case, non-linguistic factors such as personality, rapport with students and concern about students' learning may play important roles. Therefore, the more non-language abilities that the ITA test engaged and the more influence that the non-language components had on the overall evaluation of the ITA test performance, the weaker the relationship was between the TOEFL iBT Speaking Section scores and the ITA test scores.

As it requires a minimal threshold language level for communication strategies to aid communication, the TOEFL iBT Speaking Section, as a test of academic speaking skills, may be an effective measure to screen high level students who are well qualified for teaching and really low level students whose language abilities are below the minimal threshold level. Therefore, it is appropriate to use the TOEFL iBT Speaking Section as an initial pre-arrival screening measure. For borderline students, authentic performance-based tests that require language use in simulated instructional settings may help us to better assess their oral communication skills and their readiness for teaching assignments.

Setting Cut Scores on the TOEFL iBT Speaking Section for ITA Screening

It was found that the TOEFL iBT Speaking Section scores were generally accurate in classifying students into distinct TA assignment groups, the classification accuracy ranging from 71.4% to 96.7% for the model-building samples. For each school, cut scores were recommended in light of the need to minimize the chances of non-passes being classified as passes (Table 9). At UCLA and UNCC, the TOEFL iBT Speaking Section scores were also found to function reasonably well in predicting TA assignments using an independent sample with cut scores determined via the model-building sample.

Table 8: Tasks and scoring rubrics of the local ITA tests

	SPEAK	UCLA TOP test	UNCC Presentation tests	Drexel DIP test	UF Teach Evaluation
Tasks	Semi-direct test on topics of general or intellectual interest	*Simulated* teaching test (content and non-content combined)	*Simulated* teaching test (separate content and non-content based tests)	*Simulated* teaching test (content-based)	*Real* classroom teaching sessions
Rubrics	Linguistic	Linguistic	Linguistic Teaching	Linguistic Teacher presence & nonverbal communication	Linguistic Lecturing Cultural /Teaching

Conclusion

This study has shown moderately strong relationships between TOEFL iBT Speaking Section scores and local ITA test scores. It has also provided an example of how cut scores can be derived when examinees' performance levels on criterion measures are available. The results have considerable potential value in providing guidance on using the TOEFL iBT Speaking Section scores for ITA screening purposes.

It has to be noted that a recommended cut score for one school being higher than that of another does not necessarily suggest that the former requires stronger speaking skills for their TAs than the latter. The presence of a particular type of student in a sample from a particular

Table 9 Summary of the TOEFL iBT Speaking Section Cut Score Recommendations at the Four Institutions

	Pass	Provisional Pass	Criterion measure	Cross validation
UCLA	27	23-24	In-house teaching simulation test	Yes
UNCC	24	Not available[2]	In-house teaching simulation test	Yes
Drexel	23[1]	Not available[3]	ITA course instructor recommendation	No
UF	27-28	23	SPEAK	No

1. For unrestricted teaching assignments, including those requiring large-group instructional contact.
2. At UNCC, students are classified as either pass or fail based on their scores on the local ITA tests.
3. At Drexel, students may be classified into three categories: no instructional contact (NC), restricted assignments (RA), or non-restricted (all) assignments (AA). However, because none of the participants in this study was classified as an NC, it was not possible to establish a cut score for the restricted assignments (RA).

institution that did not fit the general prediction model may push up the cut score for that institution as part of the process of minimizing false positives. Thus, an institution needs to think carefully about the characteristics of their ITA population and the kind of language support available before establishing their cut scores.

The TOEFL iBT Speaking Section score recommendations for the four institutions were derived based on the participant samples used in this study. These cut scores need to be closely monitored, validated with new samples in local settings if possible, and modified if necessary. Mechanisms should be established to rectify cases where ITA assignment classification is not accurate.

Another point worth mentioning is that in this study, the consequences of having potentially unqualified ITAs (false positives) was considered more severe than those of excluding otherwise qualified ITAs (false negatives). Depending on the situation of a particular school, the ITA program may be willing to bear the consequences of having a slightly higher false positive rate to reduce the chances of classifying qualified ITAs as unqualified based on their TOEFL iBT Speaking Section scores. This is certainly a legitimate approach, assuming a mechanism could be established to rectify the situation when unqualified ITAs are put into the classroom, such as setting up a procedure to identify them and then to provide them with the language support they need.

References

Agresti, A. (2002). *Categorial data analysis* (2nd ed.). New York: Wiley.

Hoejke, B., & K. Linnell. (1994). 'Authenticity' in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, *28*, 103-125.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression (2nd ed.).* New York: Wiley.

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