June 2006
Volume 10, Number 1

Standard setting for next generation TOEFL Academic Speaking Test (TAST): Reflections on the ETS Panel of International Teaching Assistant Developers[1]

Dean Papajohn
Pima County, Tucson, Arizona, USA
<Dean.Papajohn@dot.pima.gov>

Abstract

While many institutions have utilized TOEFL scores for international admissions for many years, a speaking section has never before been a required part of TOEFL until the development of the iBT/Next Generation TOEFL. So institutions will need to determine how to set standards for the speaking section of TOEFL, also known as TOEFL Academic Speaking Test (TAST). International Teaching Assistant (ITA) developers as a group have a long history of assessing oral English through semi-direct tests, interviews, and performance tests (Ginther, April, 2004; Briggs et al., 1997; Briggs, 1994; Smith et al., 1992; Plakans & Abraham, 1990). Educational Testing Service (ETS) convened a panel of nineteen ITA developers in Philadelphia on Sept. 24, 2004. This article reflects on that panel specifically and about standard setting in general. Although the focus here is on the speaking component of TOEFL, the process for standard setting recommended by ETS for the other three sections of TOEFL, namely writing, listening, and reading, is similar. Resulting cut off scores are only as good as the standard setting process from which they are derived. Understanding the standard setting process will lead to a more accurate interpretation of standards.

TOEFL Academic Speaking Test (TAST)

TAST is an exam of oral English ability in academic contexts for nonnative speakers of English. The six test items take approximately 20 minutes to complete, for a maximum speaking time of 5.5 minutes. The first two test items are considered independent, that is the spoken response is not dependent on either reading or listening ability since the questions are both written and spoken. The next two items integrate reading and listening with a spoken response expected. In other words, the examinee is provided with some information through reading and some additional information through listening. Based on these two sources of information the examinee constructs a spoken response. The final two items integrate listening with speaking. That is, the examinee listens to a brief dialogue or a brief lecture on which to base a spoken response. Each test item is rated by a different rater on a scale from 0 to 4. The final score is an average of the six item scores and is then scaled on a 0 to 30 point scale. In comparison to the Test of Spoken English (TSE), the TAST endeavors to emphasize academic contexts and authentic academic discourse (Biber et al., 2004).

Standard setting process

There are many standard setting and adjusting processes; 38 are described by Berk (1986). The ETS standard setting assumes speaking competency is a continuum rather than a state, and decisions follow a judgment-emperical process (as defined by Berk, 1986). In a judgmental-empirical process the judges' decisions are based on performance data. Standard setting is not a task for the meek, or as Berk (1986, p. 137) states, "The process of setting performance standards is open to constant criticism É and remains controversial to discuss, difficult to execute, and almost impossible to defend." Given this environment, it is unusual that ETS provides no rationale for its chosen method of standard setting. One possibility is that ETS was looking for methods that had been previously tried and a process that could be understood by laypeople.

The day of the ETS standard setting panel the process was divided into seven steps, namely:

1. Review TAST items and scoring rubric
2. Review ITA tasks
3. Develop concept of "minimally acceptable speaker"
4. Listen to responses
5. Make preliminary recommendations
6. Discuss recommendations
7. Make final recommendations

Each of these steps will be briefly discussed.

Step 1: Review of TAST items and scoring rubric

In general, topics of a personal, campus, and academic nature lend authenticity to TAST. For assessing the ability to communicate in academic settings, the functions of explaining and summarizing as used in TAST may be more relevant than giving directions on a map or sharing a complaint or apology, as on the old TSE (Test of Spoken English) or SPEAK (Speaking Proficiency English Assessment Kit). However, there may be other relevant functions that are missing from TAST. For example, international teaching assistants may find themselves asking questions, probing for information, giving instructions for an assignment, etc. Some test items may be harder or easier than others (Faggen, 1994). For example; independent tasks may be easier for some examinees (Steele & Papajohn, 2005). While some standard setting methods introduce weighting, the standard setting process recommended by ETS did not.

Integrated tasks may confound speaking with listening and reading abilities, which may make the assessment more realistic but not necessarily more accurate. The scoring rubric attempts to focus on important features of communication. However, some important abilities do not seem to fall neatly into the rubric. For example, the *distribution of new information* (that is, how compact or diffuse a response is) may influence teaching effectiveness. Information that is communicated in a way that is too compact may be difficult to comprehend. On the other hand, information that is too diffuse may make it more difficult to understand the connections. Perhaps response lengths of 60 seconds do not allow for accurate evaluation of compactness or diffuseness. Another feature not explicit in the rubric is *strategic competence* (which includes assessing the language situation, planning what to say, and executing the communication). International teaching assistants could use strategic competence when checking for student comprehension, or rephrasing for student understanding, or for clarification of a student comment or question. At this time it is unclear how distribution of information and strategic competence are incorporated into the TAST rubric.

Step 2: Review of ITA tasks

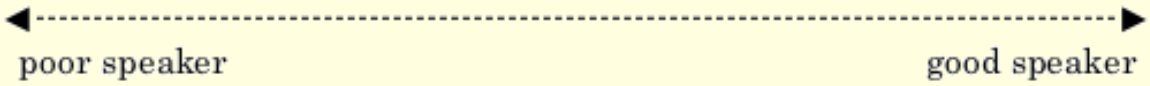
ITAs can be assigned to a wide range of responsibilities. Key ITA discourse categories, adapted from Axelson and Madden (1994) include:

- Establishing the learning environment
- Facilitating collaborative learning
- Facilitating student performance or problem solving
- Explaining content
- Questioning
- Providing feedback on assignments
- Managing the classroom
- Interacting with other instructors

It is nearly impossible to agree on the typical ITA role, which in turn makes it extremely difficult to agree on standard setting. Relevant questions include: Do all ITAs, graders, lecturers, lab instructors, and discussion leaders, need the same language skills? Do the tasks on TAST reflect important language skills for ITAs? And, what level is needed in these various language skills? Results from the panel were not based on the panel members coming to a consensus on these questions. Even if panelists recommend the same cut off score, that does not guarantee they had the same conception of the tasks expected for ITAs at their own institutions.

Step 3: Concept of "minimally acceptable speaker"

The concept of minimally acceptable speaker is judgment-based. It is intended to represent an international student whose English is just good enough to teach, or as Faggen (1994, p. 3) puts it "Éthe threshold which separates those who have sufficient content knowledge and skills." While Faggen's remarks refer to the assessment of teachers, for speakers of a second language content knowledge could refer to knowledge of the target language, and skills could be the ability to use this knowledge to communicate. Defining minimally acceptable speaker is a challenge for a standard setting panel composed of different people with different ways of thinking about communicating and teaching. The descriptor of minimal may be interpreted as a negative quality by some (as in the idea of incomplete), or as a positive quality by others (as in the idea of effective). Therefore, minimal may not be the best descriptor to use for standard setting.



Based on your own experience, would you place a minimally acceptable speaker closer to poor speaker, or closer to good speaker? It is difficult to obtain agreement on an exact placement. Competence has been described as a continuous variable and as such "there is clearly no point on the continuum that would separate students into the competent and the incompetent" (Jaeger, 1989, p. 492). If it is desirable to have high accuracy and minimal problems, then do we want the acceptable score to be defined by the word "minimal?" What abilities should an ITA have? No one communicates perfectly 100% of the time, so: How often will an ITA have to repeat him or herself to be considered minimally acceptable? What percent of undergrad students will easily understand a minimally acceptable ITA? And, how adept at interaction should a minimally acceptable ITA be? Finding agreement is encouraged by ETS, who instructs standard setting panels to:

> Acknowledge that there are differences in demands for speaking skills across disciplines and try to emphasize common speaking demands that cut across disciplines. The goal at the end of the discussion is to have reasonable consensus and understanding of panelists' expectations/standards for minimally acceptable speaking skills. (2004, Tab 3, p. 31)

However, finding agreement on the notion of minimally acceptable speaker can be elusive for a panel coming from numerous academic institutions. It is possible that different sets of judges would end up

with different results. Jaeger (1976, p. 23) states, "Given the same body of information on the nature of the domain of tasks, it is likely that different samples of judges Éwould set somewhat different standards." The panel ETS formulated had representatives from large and small institutions, public and private, and intensive English programs and teaching centers. Yet the ETS executive summary (ETS, 2005) for this standard setting session does not describe how panel members were chosen. Was the panel a representative sample of institutions with ITA programs in the United States? Panel discussions of ITA tasks and minimally competent speakers, and iterations of independent judgments of sample test performances followed by panel discussions can help alleviate bias; however, individual backgrounds can still influence judgments (Papajohn, 2002). To better interpret the results of a panel it is important to know who the panelists were and why they were chosen, and how conceptions of minimally acceptable may vary for different people and institutions.

Step 4: Listen to responses

At the ETS panel of ITA developers, some panelists thought some responses that were given a rating of 3 matched the score descriptors for level 2. In general, a level 3 response is appropriate for the task but may lack some content development, with only occasional problems with fluency. In general, a level 2 response is appropriate for the task but has limited content development, with problems in pronunciation, coherence, or fluency that interfere with communication. After much discussion about whether some of the responses rated at level 3 were accurately benchmarked, this issue remained unresolved. Since recommendations are only as good as the input they are based on, this question deserves additional attention in order to better interpret results from the panel.

Another question raised was whether independent tasks would be considered easier and thus raters will have higher expectations, or will raters think independent tasks are easier and not be as critical in their ratings. Another concern was that some panelists felt the benchmark responses did not represent the normal breadth of abilities of ITAs. Without the highly proficient speakers in the sample, the rest of the speakers may be perceived as sounding better than they actually do. Additionally, after hearing a few responses to the same questions, will listeners come to expect certain kinds of responses and be less open to other types of responses? Some panelists also wondered about the order in which sample responses were presented to the panel. Currently the samples are played from low scores to high scores. Knowing the scores may bias a cut score recommendation toward a preconceived notion of a reasonable cut score. That is, some panelists may be operating on what Faggen (1996) calls a benchmark method which focuses on score level descriptors, whereas others are operating on a test level pass/fail method which focus on examinee responses. The effect of revealing scores to panelists is uncertain, as Hambleton et al. state:

> It is not clear whether providing the panelists with the scores for examinees' papers [responses] will serve as a biasing factor early in the judgmental process, or whether the presence of scores later in the process would help alleviate random error or unwanted bias. (2000, p. 364)

Because these issues remain unresolved it makes it more difficult to interpret the results from this standard setting panel.

Steps 5 and 6: Make recommendations and discuss

The context an ITA developer works in can influence how they view standard setting. Decision makers may want to consider their own beliefs "about the effect of passing less than qualified applicants or failing qualified applicants" (Faggen, 1994, p. 6). Institutional contexts may vary on a number of characteristics, such as:

- Number of ITAs on campus
- Profile of undergraduate students on campus
- Existing cut score
- Availability of institutional support
- Source and strength of ITA policy
- Speaking level of current applicants
- Value of cultural diversity
- Anticipated future employment of ITAs

For example, if you do not have instructional development staff that can work with ITAs before or while they are teaching, you may choose a higher cut score, or if graduates from your institution generally return to their home countries to work rather than finding work in the U.S, you may recommend a lower cut score. Panelists from diverse campus settings can not help but bring such diverse perspectives to the standard setting process which can make finding consensus difficult. In fact, Faggen (1994, p. 4) contends, "Consensus is not the goal of these iterations." Using actual examinee performances as the basis of discussion, as is done in the standard setting process used by ETS, can help alleviate the pressures of social comparison of the panelists (Berk, 1986).

Step 7: TAST cut score from panel

Means, medians, highs, lows, and standard deviations were calculated for the score cut offs compiled from the panelists. The raw scores for each examinee were converted to a scale of 30. Although in reality individual examinees may score differently on various test items, the benchmark tapes represent examinees who are benchmarked at the same score level for all item responses used as samples in standard setting. This may eliminate the more difficult-to-rate examinees from the pool of responses used in standard setting, those examinees whose responses fluctuate between score bands. For ITAs, the mean score of the panelists suggests a TAST score of 23 for the cut off for minimally acceptable speaker. It appears that the cut off was based on the majority. But what does that mean for panelists that rated higher or lower than the majority? Communication is a two-way street, so different listeners are bringing different skills and background experiences to the comprehension process. No discussion took place of an appropriate percentage of students finding the ITA understandable. For example, at one cut off score 70% of undergraduate students may find ITAs comprehensible, at another cut off score 80% of undergraduate students may find ITAs comprehensible. Institutions may find different comprehensibility rates acceptable. This type of discussion was missing from the standard setting process. In addition to the cut off score recommendation for ITAs, results from the standard setting panel also suggest that a TAST score of 26 is equivalent to a score of 50 on the TSE or SPEAK.

Conclusions

Based on these observations it should be clear that the results from the ETS panel of ITA developers should be looked at cautiously and interpreted carefully. "All standard setting is judgmental," as Jaeger (1976, p. 22) succinctly puts it. So, what can we do with the results from the ETS sponsored standard setting panel? The point of this article is not to criticize the standard setting process recommended by ETS. They have applied a number of methods used in assessment and research in the past. However, it is important to know why certain methods were chosen, what the strengths and weakness of different methods are, and the issues that are raised in the steps of standard setting. Understanding the standard setting process can lead to a better interpretation of the standards that are set.

The results from the ETS standard setting panel should be considered as a starting point rather than as an end in itself. It is data you can use as a comparison to your own institutional context. Before establishing a cut off score for your institution you should investigate ITA responsibilities, go through local standard setting, and follow through with validating the cut score. The contrasting groups method could be applied at the campus level. ESL instructors could be asked to classify their students as ITA-

ready or not, then the overlap of frequency distributions of scores could be reviewed (Hambleton et al., 2000). In addition, you can also utilize other assessments such as, listening cut scores, interview, and performance tests. ITA supervisors and students of ITAs can be surveyed to validate the cut off score for ITAs (Papajohn, in press). Finally, rather than adopt or accept someone else's cut score, use the standard setting process as an opportunity to raise awareness of ITA issues at your institution.

Note

[1] Based on a presentation given at TESOL in San Antonio, March 31, 2005

References

Axelson, E. R., & Madden, C. G. (1994). Discourse strategies for ITAs across instructional contexts. In C. G. Madden& C. L. Myers, (Eds.), *Discourse and performance of international teaching assistants* (pp. 153-185). Alexandria, VA: TESOL.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Education Research, 56(1),* 137-172.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph Series MS-25). Princeton, NJ: Educational Testing Service.

Briggs, Sarah L (1994). Using performance assessment methods to screen ITAs. In C. G. Madden & C. L. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 63-80). Alexandria, VA: TESOL.

Briggs, S., Clark, V., Madden, C., Beal, R., Hyon, S., Aldridge, P. et al. (1997). *The international teaching assistant: An annotated critical bibliography* (second edition). Ann Arbor, MI: The English Language Institute, The University of Michigan.

Educational Testing Service (2004*). iBT/Next generation TOEFL standard setting: Facilitator notebook*. Princeton, NJ.

Educational Testing Service (2005, January 3*)*. Executive summary: ITA standard setting session for the Next Generation TOEFL. Message posted to the ITA Interest Section Listserve of TESOL.

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

Ginther, A. (2004). International teaching assistant testing: Policies and methods. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (2nd ed.) (pp. 57-84), Washington, D.C.: NAFSA.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (200). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24(4),* 355-366.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (pp. 485-514). NY: American Council on Education and Macmillan Publishing.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. *Florida Journal*

of Educational Research, 18, 22-27.

Papajohn, D. (2002). The standard-setting process for the new Test of Spoken English: A university case study. In W. Davis, J. Smith, & R. Smith (Eds.), *Ready to teach: Graduate teaching assistants prepare for today and for tomorrow* (pp. 167-176), Stillwater, OK: New Forums Press.

Papajohn, D. (in press). Student perceptions of the comprehensibility of international instructors. *Journal on Excellence in College Teaching*, Oxford, Ohio: Miami University.

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.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198). Baltimore: The John Hopkins University Press.

Smith, R. M. Byrd, P., Nelson, G. L., Barrett, R. P., & Constantinides, J. C. (1992). *Crossing pedagogical oceans: International teaching assistants in U.S. undergraduate education* (ASHE-ERIC Higher Education Report No. 8). Washington, D.C.: The George Washington University, School of Education and Human Development.

Steele, D., & Papajohn, D. (2005, March*). Piloting the speaking section of the Next Generation TOEFL*. Paper presented at the 39th Annual TESOL Convention, San Antonio, TX.

About the Author

Dean Papajohn has worked with international teaching assistants for over twelve years teaching English as a second language classes, coordinating orientations for international teaching assistants, and conducting oral English assessment. The second edition of his book *Toward Speaking Excellence* has been released by the University of Michigan Press.

Editor's Note: The HTML version contains no page numbers. Please use the PDF version of this article for citations.