Doing Standard Setting for an In-House High-Stakes Speaking Performance Test

Greta Gorsuch *
Texas Tech University

Abstract

Standard setting is the means by which we connect test scores to learner performance. This is important, as we need standard setting in order to set cut scores for high-stakes tests. We typically associate standard setting with large testing companies. In contrast, this is a report from an educational program in the U.S. that works with high-intermediate English learners on their speaking skills. The learners were being assessed as college-level teachers in science, math, and the humanities. An eight-hour standard setting procedure on the teaching simulation test used in the program was designed and carried out using seven video-recorded performances. This resulted in descriptions of passing and failing learners, a cut score recommendation, and rater training materials. A step-by-step generalizable procedure for a standard setting procedure is described, as well as accepted ways of evaluating the procedure.

Introduction

Standard setting is the means by which teachers, administrators, and other stakeholders connect test scores to learners' performances, as opposed to arbitrarily setting test cut scores of "pass" or "fail" at somewhere around a 60% mark. Standard setting is the process of setting cut scores using careful deliberation and accepted methods (Gorsuch & Griffee, 2018). For an accessible history of standard setting and evolving trends and methods focused on changing insights on psychological measurement, and diversifying audiences for standard setting, see Cizek (2001). There are many accepted methods, including multiple variations of the Contrasting Group Method, and the Angoff Method. The two are mentioned here because they illustrate effectively the kind of thinking behind standard setting for different types of tests. The Contrasting Group Method, and variations, involves having a panel of experts examine samples of learners' writing or speaking, defining which work samples constitute passing and failing performances, and then lining up the passing and failing performances with the scores that have been given to the learners by trained raters who are not on the panel. In other words, the expert panel does not know the test scores for the work samples beforehand. Their only job is to identify and describe passing and failing performances. The cut score is set at the point on the test scale at which panelists agree learners are minimally passing (Brown, 2005; Cizek & Bunch, 2007; Griffee & Gevara, 2011). The Angoff methods and its variants involve an expert panel examining test items or tasks, and then hypothesizing which items or tasks a "minimally competent" learner would likely get right. Adding up the number of items or tasks minimally passing learners would get

Gorsuch, G. (2020). Doing standard setting for an in-house high-stakes speaking performance test. *TESOL Working Paper Series, 18*, 58-74.
Website: Hawaii Pacific University http://www.hpu.edu
*Email: greta.gorsuch@ttu.edu. Address: Texas Tech University, Classical and Modern Languages and Literatures, Lubbock, Texas, 79409, USA

right, results in a cut score (Alderson, Clapham, & Wall, 1995; Cizek & Bunch, 2007). In this context, then, standard setting is a way we make test scores meaningful through expert knowledge, deliberation, discussion, and consensus. The results of standard setting are increased accountability and fairness for decisions we make based on learners' test scores, and greater transparency for language learners, teachers, and testers as to what constitutes different levels of language ability or performance on a given test (Kantarcioglu & Papageorgiou, 2011). For the current report, all of these results were sought.

Typically, standard setting is associated with large-scale tests, such as the *TOEFL* test (Wylie & Tannenbaum, 2006), the *TOEFL Junior Test* (Baron & Papageorgiou, 2016), the *Michigan English Test* (Chapman & Papageorgiou, 2010), and the *Texas English Language Proficiency Assessment System* reading test (Texas Education Agency, 2013). Testing companies and state education agencies have the financial resources to carry out standard setting projects, with panelists being selected and paid for their work; and the coordinator sending out samples of work or test items before a standard setting session, and then arranging a meeting in a central location for one or even two days with food, lodging, and transportation for the panelists (Raymond & Reid, 2008). Testing companies also have the expertise to design and carry out standard setting procedures, including hiring and training panelists (Raymond & Reid, 2008). This leaves independent programs with limited budgets in a difficult position. In this case, a grant was applied for and awarded, which made the standard setting procedure and this resulting report possible.

This report details a standard setting procedure for a high-stakes teaching simulation performance test (The International Teaching Assistant Performance Test; Gorsuch, 2006; Gorsuch, Florence, & Griffee, 2016; Gorsuch & Griffee, 2016, July). The locally developed test is found in the Appendix, and the training descriptors, which amount to twenty pages, are available upon request. The test can also be found in the IRIS database of instruments for research into second languages (https://www.iris-database.org/iris/app/home/index). This standard setting procedure (a Contrasting Groups procedure) can be adapted to performance tests used in other programs, such as English-medium universities in international settings where they seek to hire and develop a pool of English-speaking content teachers, or schools with a graduation requirement for writing ability. As long as there are reviewable samples of learner performances, such as video-recordings or writing samples, the standard setting procedure described here should work. Data presentation visuals used in the standard setting session are offered.

Background

The Program and The Test

The Graduate ESL/ITA program at Texas Tech University is the longest operating program in the U.S. and is recognized at an international level. Following global trends, many of the program's Tech graduate students are international teaching assistants (ITAs). The program works with 160+ newly arriving international scientists and graduate students per year, and use principled assessments and second language learning theories to support ITAs to teach undergraduates.

International teaching assistants (ITAs) are high-intermediate to low-advanced English speakers who are being supported as graduate students in North American universities. How are ITAs described here as high-intermediate to low-advanced? One means of description comes from the IELTS test. Some ITAs use IELTS test scores for admission at the university in question. The minimum score at the institution is 6.5 overall, which is considered between a "competent user" and a "good user" (IELTS, 2017).[1] Nonetheless, many ITAs come from countries where English is not widely used. Universities want them as graduate students because they are gifted young scientists and scholars in physics, chemistry, engineering, math, history, art, family counseling, and tourism industries. The condition of their support at universities is that they teach basic courses within their content areas to 18-25 year old undergraduate students (Gorsuch, 2016). Many ITAs are beginning-career scientists and scholars, and thus have little to no teaching experience in their first languages, much less in their second language (English). Many ITA programs use a battery of tests, including performance tests, to determine whether ITA candidates are approved to teach. These performance tests often take the form of teaching simulations. Such tests are high stakes. Thus, standard setting should be done, to ensure that valid decisions are made with the test scores, and to ensure transparency for stakeholders.

The ITA Performance Test has been used in some form since 2001, and in 2016, the test underwent a major revision inspired by continued problems with rater training. The problems stemmed from raters' continued confusion over the test criteria and descriptors for levels on each criterion (Gevara, Gorsuch, Almekdash, & Jiang, 2015). This resulted in inter-rater reliability estimates lower than .80. What was most lacking were work samples of learners' teaching simulations that could be matched to the level descriptors on criteria. This was one of the reasons for doing the standard setting project. The ITA Performance Test revision resulted in ten criteria with four categories for each: Pre-functional, Beginner, Intermediate, and Sustainably Fluent. The four categories represent the full range of ability levels seen in eighteen years in incoming ITAs. Further, the four categories are needed for diagnostic and instructional purposes. Learners get 1 – 4 points on each of the ten criteria for a total possible score of 40 (Pre-functional on a given criterion = 1, Beginner = 2, Intermediate = 3, Sustainably fluent = 4). The ten criteria are: word level pronunciation, word stress, thought groups, grammatical structures, transitional phrases, definitions and examples, prominence, audience non-comprehension awareness, tone choices, and handling questions (see the Appendix). Detailed training descriptors for each criterion and category were also written, piloted, and revised. A test procedure was documented, including suggestions to ensure each learner speaks for ten minutes on a concept or process from their field, and answers audience questions about it. A generalizable rater training procedure was designed and piloted (Gorsuch et al, 2016). In this context, "generalizable" means that other schools and users of this test, or similar tests, can use the procedure.

The Need for a Replicable Standard Setting Session for the Test

Performance testing (tests which result in a written or spoken performance and are subjectively scored) is part of a growing trend in second and foreign language programs world-wide. Creating a locally replicable standard setting procedure will benefit foreign or second language educators

or program leaders who use subjectively scored performance tests that are relevant to their local contexts.

Standard setting of the kind used here (a Performance Profile procedure with a final component using a Contrasting Groups strategy) involved choosing video-recordings of learners' teaching simulation presentations that represented the four level categories on the ITA performance test. This method was most suitable in that the assessment was a performance test with complex scoring criteria and multiple levels for each criterion, with the additional assumption that there were interdependencies between performance criteria (Hambleton, Jaeger, Plake, & Mills, 2000). Other methods, such as the Angoff Method and its variants, are more suited to examination of dichotomously scored test items. See additional notes in step 2 of the procedure for this report below. The rater training materials that would result from the standard setting included:

- Multiple video files matched with panelists' categorizations of Pre-functional, Beginner, Intermediate, and Sustainably fluent performances
- Panelists' descriptions of passing or failing performances in terms of ITA Performance Test criteria

The Standard Setting Procedure

Figure 1

Sequential model of a performance profile and contrasting groups standard setting for a performance test

1. Choose a panel; Set date and reserve room	2. Choose a standard setting method	3. Prepare work samples: Locate and get FERPA permission to use video files; Write IRB proposal for expert panel data	4. Prepare pre-session training materials (instructions, video-recorded work samples) and send out three weeks in advance	5. Finalize the standard setting session agenda	6. Get first round test ratings by physically picking up rating forms from panelists; One week later ask panelists to put learners into level placements, panelists send placement data the night before the session; Prepare seven data presentation visuals; Panelists complete a self-assessment of confidence for their placement and rating of work samples	7. Begin session, give quiz on descriptors for ability levels (used for work sample placements); Examine first round data as a group
8. Discuss first round Beginner, Intermediat e and Sustainably fluent level placements	9. Concur on one minimally passing learner, one borderline failing learner, plus any other learners with less consensus on level placement	10. Do a second round of level placements and ratings for selected work samples	11. Display second round placements and discuss; Confirm work sample for minimally passing learner and borderline failing learner	12. Write descriptions of passing (Sustainably fluent) and failing (Intermediate or Beginner) learners based on test criteria	13. Panelists evaluate pre-session training materials and standard setting session	14. Panelists rank learners 1 – 7 in descending order; Calculate median total scores for learners and line up against rankings, and obtain cut score

Because the purpose here is to offer a replicable model for readers to use, the report appears in terms of the steps of a sequential model. The information readers would normally see under conventional headings is of course in the report, but within sequenced steps of the model. For

instance, a description of the participants appears in step 1 ("Choose a panel"). See Figure 1. The steps were adapted from general assessment sources (Cizek & Bunch, 2007; Hambleton, 2000, 2008; Hambleton et al., 2000), as well as language testing sources (Baron & Papageorgiou, 2016; Tannenbaum & Cho, 2014). There was a good deal of convergence on the procedure steps between sources, with the language testing sources citing general assessment sources, also named here, in common. The motivation for each step is defined, and a description of what happened in the standard setting session is given. The descriptions serve to reveal what issues came up that were specific to language testing, and the ITA Performance Test itself, and how they got resolved. Standard setting, even based on a model, is not seamless, nor necessarily straightforward.

In the interests of space, not all of the components will be commented on. For #5, for example, the four page working agenda on which the face-to-face session was based, is available upon request.

Step 1, *Selecting qualified panelists*, and enough of them, is important, to bring diverse viewpoints and at the same time yield stable and reliable data on placement and ratings (Tannenbaum & Cho, 2014). Six panelists were found who had experience with the learners and with the test, and had at least an M.A. in applied linguistics, or were late-career M.A. students who were supported as T.A.s tutoring the learner population. Thus, they had homogeneous qualifications relative to their roles as panelists. The author made a seventh panelist. The panelists were coded 1, 2, 3, 4, 5, 6, and 7. In all data presentation visuals used for group discussion, the panelists were identified only by their code numbers. There were eight hours of paid work, four hours of which were done at home, and four hours of which were the face-to-face session.

In *Step 2*, there are many standard setting methods (for example, Hambleton, Jaeger, Plake, & Mills, 2000). Which method to choose partly depends on the type of test. For performance tests, one method called "performance profile" focuses on panelists viewing work samples of learners and "identify[ing] the profile…most likely to be earned by a test taker meeting the minimum expectation of a framework level" (Tannenbaum & Cho, 2014). This method, with elements of the contrasting groups method, was used. The procedure focused first on placements of learners into the four performance categories resulting in multiple video files matched by consensus to each of the performance categories of Pre-functional (failing), Beginner (failing), Intermediate (failing), and Sustainably fluent (passing). The second focus was on the learners' total scores on the test, reached by consensus, and how the numerical scores plotted, from high to low, against the placement of learners into performance categories (the contrasting groups of "pass" or "fail"). There is more information on how the panelists' work was sequenced to accomplish this in Figure 1 and in the descriptions of points #6 – 12 below). Because there were few qualified panelists in the region, it was not possible to have large numbers of video-recorded teaching simulations scored in advance by raters who were not part of the standard setting.

In *Step 3,* selecting the work samples required that the number of samples (video-recordings of past ITA candidates) be large enough to represent the four categories of performance and yet not too large as to be unworkable for a four-hour standard setting session

(see Burrows, Bingham, & Brailovsky, 1999 for commentary on sample sizes and standard setting in small programs). Twenty-two video files were selected that represented the range of learners from Pre-functional to Sustainably fluent. According to federal regulations (Family Educational Rights and Privacy Act, U.S. Department of Education, 2015) and the university counsel, each learner had to be contacted to ask permission to use their video recording for the standard setting. Two waves of e-mails with the permission form were sent out, and seven of the learners agreed to release their video for use. Ultimately, the learners' video files represented Beginner, Intermediate, and Sustainably fluent levels, but not the Pre-functional level. The seven video files represented candidates from India, Sri Lanka, Vietnam, China, and Korea.

Step 4 is a key component of the panelist training (Cizek & Bunch, 2007; Raymond & Reid, 2008), where the pre-session packet was sent out three weeks before the standard setting session. It was 15 pages long, and had instructions, self-quizzes, training notes on standard setting, and rating and category/level placement forms. This is an excerpt from the training notes:

> First, as a program, we need to confirm or disconfirm our current cut score. By focusing on the two operations of rating and categorizing separately, and without conscious reference to each other, we can discover where there is a consensus on a cut score. If there is "softness" or "imprecision" on cut scores, we can use discussion to write descriptions of these "borderline" cases and make decisions about how to pass or fail in borderline cases.
>
> Second, we can study the relationship between passing and failing students and their scores on individual ITA Performance Test criteria.
>
> To reiterate, the logic of this method stipulates first rating the seven video files on the full test, and then setting those ratings aside for three or more days. Then the video files are viewed again and classified into "pass" or "non-pass." Finally, during the actual session on October 21, we then relate the scores (the ratings) to passing or failing classification.
>
> **Self-evaluation quiz**. Please order the steps. Please do not look at the previous page until you have ordered the steps. You can re-order them after checking your answers:
>
> _____ We discuss single criterion scores with reference to passing or failing classification
>
> _____ We classify test candidates into passing or non-passing groups without reference to our ratings
>
> _____ We put total test scores into a distribution and compare them to our judgments of passing or failing classifications
>
> _____ We rate the videos on the ITA Performance Test

The seven video files were also included on a DVD. The video files were coded by color from the spectrum and the topic of the presentation: "Red Set and Function" (a mathematician), "Orange Genes" (a biologist), "Yellow First Hand Account" (an historian), "Green Flame Reaction" (a

chemist), "Blue Sensitivity Analysis" (an engineer), "Indigo Factors of Production" (an economist), and "Violet Food Chain" (a biologist).

In *Step 5*, the session agenda was finalized and handed out 24 hours before the session.

In *Step 6*, in the first week of the pre-session process, panelists rated the seven learners' video files using the ten criteria of the ITA Performance Test (see the Appendix). The author picked up the forms from panelists' homes or workplaces. Panelists were then asked to spend one week away from the video files. The author imposed the same injunction upon herself. Then, just before the session, the panelists re-viewed the seven video files and categorized each performance into one of the four overall performance categories of Pre-functional (failing), Beginner (failing), Intermediate (failing), and Sustainably fluent (passing). It was important that panelists work with the categorizations without reference to their ratings of the week before, so as to more effectively elicit thinking about which learners passed and why, and which learners failed and why. Rating the video-recordings and categorizing the learners into the four categories had to be different operations. One week away from the data helped ensure that would happen.

In *Step 7*, the resulting seven visuals (one for each learner) used as the focus of discussion can be seen in Figure 2 for "Red Set and Function." The categorizations were shown for each rater (1 through 7). The scores came from the ratings the panelists had done the week before. See Reckase (2008) for alternate visual tools.

Figure 2
Data presentation visual for candidate "Red Set and Function"

Panelist	Categorization I	Red set and function test score I
1	Sustainably fluent	38
2	Sustainably fluent	37
3	Sustainably fluent	38
4	Intermediate	36
5	Sustainably fluent	40
6	Intermediate	39
7	Beginner	35
Median		38
Mean		37.571
SD		1.718
Min/Max		36/40
	Sustainably fluent = 4 Intermediate = 2 Beginner = 1	

In *Step 8*, the visuals were key to providing panelists with normative feedback, which showed how their categorizations and scores compared to those of the other panelists'. It is a kind of reality check in which "participants see the extent to which they are judging more harshly or leniently relative to other participants" (Cizek & Bunch, 2007, p. 55; see also Raymond & Reid, 2008). Discussion over all seven of the candidates was lively. Here is one example: One ITA who was determined by consensus to be sustainably fluent, but minimally so, raised ensuing commentary involving nearly every test criterion. Related to the thought groups criterion, some

panelists thought the "Red Set and Function" ITA used "good thought groups" while others said he used "lots of uh….uh." One other panelist agreed that was true, but that the learner "uses fillers but they do not interfere with audience understanding." Recall that this was not a rater training session. The standard setting session was not intended to argue over the test. Rather, it was to reach consensus over the categorizations of the learners overall, using the test criteria as resources for the discussion. Panelists were encouraged to give their reasons for their categorizations and their ratings. In the case of four files where there was less agreement, the video files were reviewed repeatedly as directed by panelists and used by panelists to explain their impressions. The discussion based on the visuals proved to be a significant way for panelists to learn to describe their categorizations in terms of the test criteria. Examples are given in Figure 3.

In *Step 9,* the panelists concurred that "Red Set and Function" represented a minimally passing candidate, while "Blue Sensitivity Analysis" represented a borderline failing candidate. They concurred that two additional candidates needed more discussion, including "Yellow First Hand Account" and "Orange Genes."

This resulted in *Step 10*, which was a second round of category placements, and then ratings, of the four candidates.

These decisions resulted in *Step 11,* which was a second data presentation visual of the four candidates comparing first round and second round placements and ratings. This second visual was viewed and discussed by panelists and it was concurred that no further discussion was needed. The panelists believed the categorizations were appropriate.

In *Step 12,* the panelists then wrote notes and discussed descriptions of performances that contributed to learners passing or failing (Sustainably fluent versus Intermediate or Beginner). The second visual, with the descriptions, again for "Red Set and Function," is found in Figure 3.

For all four of the candidates whose video files were reviewed during the second round, there was closer agreement between the panelists, as evidenced by smaller standard deviations (1.718 versus .976 in Figure 3 for "Red Set and Function")[2] and greater consensus on overall descriptor categories (Sustainably fluent = 4, Intermediate = 2, Basic = 1 for the first round of ratings; Sustainably fluent = 5, Intermediate = 2 for the second round of ratings, see Figure 3). This suggests that the standard setting procedure (Figure 1) was effective (Hambleton, 1999; 2008). The procedure stipulated multiple rounds of discussion, and clearly the discussion, an artifact of the procedure, brought about closer agreement. The three candidates who were not reviewed a second time already had a close consensus on scores and placements from panelists.

Figure 3

Second visual comparing first and second round ratings and placements for one candidate

Panelist	Categorization I	Categorization II	Red set and function test score I	Red set and function test score II
1	Sustainably fluent	Sustainably fluent	38	38
2	Sustainably fluent	Sustainably fluent	37	37
3	Sustainably fluent	Sustainably fluent	38	38
4	Intermediate	Sustainably fluent	36	36
5	Sustainably fluent	Sustainably fluent	40	39
6	Intermediate	Intermediate	39	38
7	Beginner	Intermediate	35	37
Median			38	38
Mean			37.571	37.571
SD			1.718	0.976
Min/Max			36/40	36/40
	Sustainably fluent = 4 Intermediate = 2 Beginner = 1	Sustainably fluent = 5 Intermediate = 2		

This candidate was determined to be **minimally passing**

Comments of strengths which contributed to the candidate passing were:
•Starts with interaction
•Asks teaching questions; *Do you understand what I mean by well defined?; Does it make sense?*
•Uses sophisticated comprehension checks such as *Do you understand what I mean by well-defined? Does it make sense?*
•Uses analogies
•Repeats key terms
•Uses good thought groups, uses fillers but do not interfere with audience understanding
•Uses some prominence
•Uses some variations in tone choices
•Handles questions well
•Uses clear examples

Comments on weaknesses which *could* contribute to the candidate failing were:
•Major word stress problems *domain*
•Seems inexperienced in presenting content
•Lots of *uh….uh*
•Searching for words
•Abandoned thoughts, dangling thoughts, so audience gets lost at pauses
•Choppy, distracting thought groups
•At 8:21 unclear how we got to the co-domain concept
•Rushes through some thought groups
•I wish he had checked for comprehension more

Note This individual was rated three times, once with raters working alone prior to the standard setting session, once again during the standard setting session, and once more during second round ratings. All comments above come from notes made by panelists at the standard setting session.

In *Step 13*, six of the panelists (excluding the author) completed an anonymous post-session evaluation, which is consonant with good practice (Hambleton, 1999; Tannenbaum & Cho, 2014). Five of six panelists felt the standard setting packet was helpful to understanding the purpose of the standard setting project, and six of six thought the explanation of the logic behind the standard setting method seemed clear. All panelists felt confident that their pass/fail and performance categorizations of the candidates were appropriate. Five of six indicated they could understand the data as presented in the session, and six of six said it helped them to see other panelists' pass/fail categorizations. Six of six felt their opinions and views were listened to, and six of six said they had an idea of how to identify a minimally passing performance.

Finally, in *Step 14*, learners were ranked 1 (highest) to 7 (lowest) by the panelists in terms of overall performance level. The median second-round scores of the four reviewed candidates and the median first-round scores of the three candidates upon whom the panelists concurred did not need review, were lined up with the category placements, and the cut score of 38 was obtained. See Burrows et al (1999) for a similar graphical representation that shows the logic behind this practice (see Table 1).

Table 1
Implicational score and categorization scale

Categorization	Median performance test score
Sustainably fluent (pass)("Indigo Factors of Production")	40
Sustainably fluent (pass)("Violet Food Chain")	40
Sustainably fluent (pass)("Green Flame Reaction")	39
Sustainably fluent (minimally passing)("Red Set and Function")	38
Intermediate (borderline failing)("Blue Sensitivity Analysis")	36
Intermediate ("Yellow Second Hand Account")	36
Beginner ("Orange Genes")	33

It is true that no test candidate received a median test of score of 37. A reader might reasonably argue that 37 could be a cut score. Yet the learner at 38 was considered "minimally passing" by the panelists. There would be little expectation that a learner getting a 37 would pass. At the same time, this is a conservative test in two senses. First, the institution has more concern over ITAs passing the ITA Performance Test, but then not having the basic capability to communicate in classrooms with U.S. undergraduates. Second, the ITA Performance Test is but one of three required tests to be approved to teach at the institution. Failure on any one would result in a test candidate not being approved to teach. It is rare for learners in the program to fail only one of the three required tests. Thus, when learners fail one test, they usually fail at least one other test.

Limitations and Suggestions

The pool of panelists was small, and the sample of video files was smaller than might be found in standard setting with large testing companies. It was difficult to find more than seven video files, given the constraints. First, there weren't many pre-functional test candidates at the institution at the time. Most current international graduate students did not fit into the category. And second, FERPA rules, and the self-imposed restriction of not doing a third request, reduced the number of video files. A suggestion for test creators in small institutions would be to pool video files, or whatever samples of work used in the standard setting (writing samples, for instance) for two to three years to ensure sufficient samples that represent all performance levels of the test in question. One other suggestion is to have two facilitators for the standard setting. The kind of knowledge developed by doing standard setting is valuable and hard-earned. Having two facilitators might reduce the workload, and ensure that standard settings are done after two or three years, or after any major change in the test or change in the learner population.

Conclusion

This report gives the basic outlines of this standard setting session in a replicable, sequential model form. Ways of evaluating the standard setting were pointed out, including quizzing panelists on the standard setting session content, and asking for their evaluation after the session (Tannenbaum & Cho, 2014). It was also shown how greater convergence in second round data could be shown in narrower standard deviations and more agreement on category placements (Baron & Papageorgiou, 2016). Despite having only seven video files and having a limited pool of panelists from which to hire, the model seemed workable. The process resulted in a new cut score, and valuable rater training materials good for at least two years.

Endnotes

[1] See Kaufman & Brownworth (2006) for case studies of many ITA programs and their learner population descriptions and testing practices.

[2] First *SD* = 3.155 versus second *SD* = 2.138 for "Orange Genes," first *SD* = 1.54 versus second *SD* = .90 for "Yellow First Hand Account," and first *SD* = 1.464 versus second *SD* = .69 for "Green Flame Reaction"

Acknowledgement

I wish to thanks Texas Tech University and their Catalyst Grant Program.

References

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Baron, P. & Papageorgiou, S. (2016). Setting language proficiency score requirements for English-as-a-second-language placement decisions in secondary education. Princeton, NJ: ETS Research Report Series: Educational Testing Service. ISSN 2330-8516.

Brown, J.D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Burrows, P., Bingham, L., & Brailovsky, C. (1999). A modified contrasting groups method used for setting the pass mark in a small scale standardized patient examination. *Advances in Health Sciences Education, 4*, 145-154.

Chapman, M., & Papageorgiou, S. (2010). Insights in language testing: An interview with Spiros Papageorgiou. *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(1), pp. 14-18.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Gevara, J.R., Gorsuch, G., Almekdash, H., Jiang, W. (2015). Native and non-native English speaking ITA performance test raters: Do they rate ITA candidates differently? In G. Gorsuch (Ed.), *Talking matters: Research on talk and communication of international teaching assistants* (pp. 313-346). Stillwater, OK: New Forums Press.

Gorsuch, G.J. (2006). Discipline-specific practica for international teaching assistants. *English for Specific Purposes, 25*(1), 90-108.

Gorsuch, G. (2016). International teaching assistants at universities: A research agenda. *Language Teaching, 49*(2), 275-290.

Gorsuch, G., & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research*. Charlotte, NC: Information Age Publishing.

Gorsuch, G. & Griffee, D. (2016, July). Ways to evaluate rater training for ITA performance tests. *International Teaching Assistant Interest Section (ITAIS) Newsletter.* Retrieved July 8, 2016 from: http://newsmanager.commpartners.com/tesolitais/issues/2016-06-22/3.html

Gorsuch, G., Florence, R. D., & Griffee, D. (2016, February 16). Evaluating and improving rater training for ITA performance tests. *International Teaching Assistant Interest Section (ITAIS) Newsletter.* Retrieved February 16, 2016, from http://newsmanager.commpartners.com/tesolitais/issues/2016-02-02/3.html

Griffee, D. T., & Gevara, J. R. (2011). Standard setting in the post-modern era for an ITA Performance Test. *Texas Papers in Foreign Language Education, 15*(1), 3-16. Retrieved from http://studentorgs.utexas.edu/flesa/TPFLE_New/Index.htm.

Hambleton, R. (1999). Setting performance standards for educational assessments and criteria for evaluating the process. *Laboratory of Psychometric and Evaluative Research Report No. 377*. Amherst, MA: University of Massachusetts. Available: http://www.nciea.org/publications/SetStandards_Hambleton99.pdf

Hambleton, R. (2008). Setting performance standards on educational assessments and criteria for evaluating the process. In Cizek, G. (Ed.), *Setting performance standards* (pp. 89-116). New York: Routledge.

Hambleton, R., Jaeger, R., Plake, B., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355-366.

IELTS (2017). How IELTS is scored. Available: https://www.ielts.org/about-the-test/how-ielts-

is-scored

Kantarcioglu, E. & Papageorgiou, S. (2001). Benchmarking and standards in language tests. In O'Sullivan, B. (Ed.), *Language testing: Theories and practices* (pp. 94-110). Houndsmills, England: Palgrave Macmillan.

Kaufman, D. & Brownworth, B. (2006). *Professional development of international teaching assistants*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

North, B. (2011). Describing language levels. In O'Sullivan, B. (Ed.), *Language testing: Theories and practices* (pp. 33-59). Houndsmills, England: Palgrave Macmillan.

Raymond, M. & Reid, J. (2008). Who made thee judge? Selecting and training participants for standard setting. In Cizek, G. (Ed.), *Setting performance standards* (pp. 119-173). New York: Routledge.

Reckase, M. (2008). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In Cizek, G. (Ed.), *Setting performance standards* (pp. 159-173). New York: Routledge.

Tannenbaum, R. & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly, 11*, 233-249.

Texas Education Agency (2013). *Texas English language proficiency assessment system (TELPAS) reading: Standards review questions and answers*. Available: http://tea.texas.gov/student.assessment/ell/telpas/

U.S. Department of Education (2015). *Family education rights and privacy act (FERPA)*. Available: http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

Wylie, C. & Tannenbaum, R. (2006). TOEFL *academic speaking test: Setting cut score for international teaching assistants.* ETS Research Memorandum, RM-06-01. Available: https://www.ets.org/research/policy_research_reports/publications/report/2006/icja

Appendix

International Teaching Assistant Performance Test V.10.2

Topic: _____

ITA Candidate Name: _____ Date: _____

LAST NAME BOLD First name upper and lower

Rater: _____ Time: _____ Room: _____

1. Word-level pronunciation

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners are largely unable to understand words and terms in the talk.	Listeners have some trouble understanding words and terms in the talk.	Listeners can understand words and terms in the talk but with effort.	Listeners can readily understand words and terms in the talk.

2. Word stress

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners cannot understand words and terms with two or more syllables.	Listeners struggle to understand words and terms with two or more syllables.	Listeners are distracted by some errors in words and terms with two or more syllables.	Listeners have occasional difficulty but words and terms with two or more syllables are usually understandable.

3. Thought groups

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners cannot extract ideas from the talk.	Listeners have trouble extracting ideas from the talk.	At times listeners are uncertain when an idea is complete and a new idea begins.	Generally listeners are not aware whether thought groups are used.

4. Grammatical structures

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners cannot extract information from the talk.	Listeners are confused by ungrammatical propositions, and stay confused.	Listeners are confused by some ungrammatical propositions but can sometimes pick up meaning as the talk proceeds.	Listeners are not confused by ungrammatical propositions.

5. Transitional phrases

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners cannot follow the logic of the talk.	Listeners cannot easily follow the logic of the talk.	Listeners experience occasional gaps in the logic of the talk.	Listeners can follow the logic of the talk.

6. Definitions and examples

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners do not hear examples or definitions.	Listeners have trouble discerning when definitions or examples are given.	Listeners recognize when an example or definition is given but may be confused by it.	Listeners find the definitions and/or examples useful to grasp an idea.

7. Prominence

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners cannot distinguish key terms or words, transitions, and/or contrasting concepts from the stream of words in the talk.	Listeners hear few key terms or words, transitions and/or contrasting concepts in the stream of the talk.	Listeners hear some key terms or words, transitions and/or contrasting concepts, but not consistently throughout the talk.	Listeners are generally clear on key terms or words, transitions and contrasting concepts used in the talk.

8. Audience non-comprehension awareness

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners are not given chances to clear up what they do not understand. Their confusion is not recognized nor addressed by the candidate.	Listeners are seldom given chances to clear up what they do not understand. Their confusion may be recognized by the candidate, but the candidate has few apparent resources for addressing it.	Listeners are sometimes given chances to clear up what they do not understand. The candidate sometimes recognizes confusion and may have some success addressing it.	Listeners have opportunities to clear up sources of confusion. The candidate readily and successfully addresses and resolves the confusion.

9. Tone choices

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners hear monotonous sounding speech and cannot extract ideas from it.	Listeners mostly hear monotonous sounding speech with many level and falling tones, and can extract few ideas.	Listeners periodically hear rising and falling tones in the talk and can extract some ideas.	Listeners hear a variety of rising and falling tones and can readily extract ideas from the talk.

10. Handling questions

1 - - - - - - - - - - - - - - - - 2- - - - - - - - - - - - - - - - -3- - - - - - - - - - - - - - - - - -4- - - - - - - - - - - - - 5

Pre-functional	Beginner	Intermediate	Sustainably communicative
Listeners' questions may be recognized as such but are not answered.	Listeners' questions are recognized but not necessarily answered.	Listeners' questions are recognized and are sometimes answered.	Listeners' questions are usually answered.

What can be improved:

_____

_____

For reference:

Overall descriptors			
Pre-functional in classroom communication	Beginner in classroom communication	Intermediate in classroom communication	Sustainably fluent and communicative in the classroom
The candidate does not use appropriate Discourse Intonation, pronunciation, or classroom communication strategies while attempting to present classroom content. While the candidate may utter some recognizable phrases or short sentences, their utterances do not effectively propose content and they are hard to follow. The candidate is likely not communicative in classrooms.	The candidate rarely uses appropriate Discourse Intonation and/or pronunciation and/or classroom communication strategies while presenting classroom content. While the candidate may be able to make a few connected content ideas apparent to classroom learners, the message is not readily coherent, and thus the candidate is likely not communicative in classrooms.	The candidate may demonstrate use of appropriate Discourse Intonation and/or pronunciation and/or classroom communication strategies while presenting classroom content, but not always at the same time, and without regularity. This is particularly true of unscripted presentation and interaction in classrooms. Thus the candidate is somewhat communicative in classrooms but not consistently so, and may unpredictably fail to exchange meaning with classroom learners.	The candidate consistently uses features of Discourse Intonation, pronunciation, and classroom communication strategies while presenting classroom content. The candidate is a reasonably effective classroom communicator and is likely to exchange meaning effectively with classroom learners.

About the Author:
Greta Gorsuch teaches applied linguistics and ESL in the U.S. She is co-author of *Second Language Course Evaluation* (2016) and *Second Language Testing for Student Evaluation and Classroom Research* (2018), both from Information Age Publishing. She is author and editor of *Tests that Second Language Teachers Make and Use* (Cambridge Scholars Publishing, 2019). Website: https://www.gretawix.com