Griffee, D. T. & Gevara, J. (2012). Analyzing item bias to validate and revise an ITA performance test. In. J. Levis & K. LeVelle (Eds.). *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference,* Sept. 2011. (pp. 195-204). Ames, IA: Iowa State University.

ANALYZING ITEM BIAS TO VALIDATE AND REVISE AN ITA PERFORMANCE TEST

Dale T. Griffee, Texas Tech University

Jeremy Gevara, Texas Tech University

Classroom teachers sometimes have an aversion to testing because they see tests as a device to fail students rather than teach them. However, when teachers are involved in a program with high stakes results, the test need to be as fair as possible. Estimating item bias is one way to evaluate a test to make it a more equitable decision-making instrument. Using the SOAC program evaluation model, this paper reports a test instrument validation study. The purpose of this study was to determine item bias on International Teaching Assistant (ITA) Performance Test version 8.3, a test designed to evaluate speech fluency and pronunciation in simulated teaching situations (Gorsuch, Meyers, Pickering, & Griffee, 2010). Using Multiple Analysis of Variance (MANOVA), we examined scores from the ten test criteria from passing and failing groups. Results showed no statistically significant difference for criterion four (ITA uses grammatical structures, word choice, and transitional phrases effectively to provide cohesion to the content) and criterion nine (ITA candidate uses visuals or multimedia effectively). Results for the other eight criteria, however, operate effectively, showing a statistical significance between the two groups.

DEFINITIONS

Bias is systematic error (Vogt, 2005), and as used here refers to the tendency of test criteria to systematically skew the test results by not performing in the way intended. *Evaluation* judges the value or worth of an educational endeavor, and *an evaluation study* gathers information in a systematic way in order to accomplish that judgment (Alderson, 1986; Brown, 1995; Lynch, 1996; Stufflebeam & Webster 1983). A *high stakes test* refers to a situation that has important consequences for test takers. A *master* is a test-taker who passes a test at a prescribed cut score and is, therefore, assumed to have mastered the material. A *non-master* is a test-taker who has not. *Validity* is an estimation of the extent to which evidence supports the interpretation of a test result (Messick, 1996). A *program evaluation model* is a working theory of how a program functions and how evaluation studies can be organized and sequenced (Stufflebeam & Shinkfield, 2007).

The SOAC evaluation model

The SOAC model, as seen in Figure 1, is a program evaluation model designed especially for second language courses, and can be helpful in explicating the role of instrument validation (Griffee & Gevara, 2011).

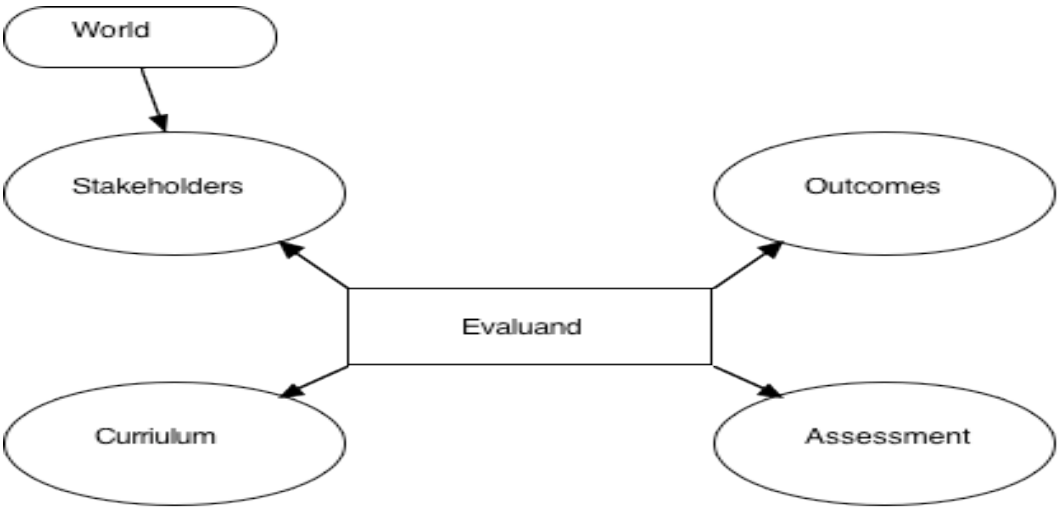

Figure 1. The SOAC (pronounced soak) model of program evaluation.

The SOAC model posits that the evaluand, the part of a program being evaluated, be evaluated in terms of four basic areas of interest: stakeholders, outcomes, assessment, and curriculum. The area of *stakeholders* includes, among other things, persons or agencies that have an interest in the outcome of the evaluation, and directly connects to the *world*, influences and pressures from outside the course. *Outcomes* include goals, objectives, or learning outcomes of the evaluand, in our case a course. *Assessment* refers to data collection instruments and other aspects of the assessment plan, and *curriculum* includes anything related to materials and teaching. The SOAC model is flexible in that any area of interest can be related to any other area and by relating the four areas of interest and the world, several evaluations areas can be identified. For example, by examining the relationship between outcomes and stakeholders, a goal validation study can be undertaken, and by examining the relationship between curriculum and outcomes, a course logic evaluation (does the curriculum logically support the outcomes) can be conceptualized. The relationship of particular interest in this study is the interaction between curriculum and assessment here called a *test instrument validation study*.

Classroom teachers and testing

Many classroom teachers enter the teaching field with little interest in test construction and validation. According to Graves (1996, p. 32), teachers feel inadequate in dealing with testing because they believe testing to be a specialist field for which they do not have adequate training. If they take a testing course, they generally find it interesting and helpful (Bailey & Brown, 1996), but the majority of classroom teachers do not take a testing class. Nevertheless, in language programs, especially those requiring high stakes decisions, tests become a relevant issue because program directors are reluctant to base high stakes decisions on a single, holistic teacher decision. Tests are valued because of their perceived potential to add a layer of objectivity and fairness to the decision-making process.

ITA programs

An international teaching assistant or ITA is a student who typically has graduated from a master's program in his or her country, and is now entering a doctoral program at a U.S. university. In return for tuition assistance and perhaps a stipend, the ITA is assigned to teach certain courses, especially first year undergraduate courses and labs. In the best case, the university department gains a high quality teacher and the ITA gains financial aid, visa support, and teaching experience (Sheridan, 1991).

Beginning in the 1980s, the number of U.S. graduate students began declining while the number of ITAs began increasing, especially in math and science (Wilkening, 1991). At the same time, increasing numbers of U. S. undergraduates were coming to college with plans for jobs upon graduation. When U.S. undergraduates met the ITAs, they sometimes complained to their parents that they could not understand the ITAs, parents complained to university administrators and state legislators, and ITA training programs were born. One such program is at our university.

The function of the present ITA program is to test incoming ITA candidates, and based on test results, to approve those who score at or above the cut scores to teach. For those ITA candidates who do not meet the cut score, remediation programs are used which include retesting. One type of test used to evaluate incoming ITA candidates is a performance test.

Based on university operating policy in response to the Texas Education Code, international students accepted into a Masters or Doctoral program and eligible to receive a Teaching Assistantship are notified of the ITA workshop, which they are required to pass. ITA candidates who do not pass have the option of taking ESL5310, a semester class equivalent to the ITA workshop. Candidates passing the course are eligible to teach the following semester.

Performance Test

A performance test (PT) requires a test candidate to do something rather than to demonstrate knowledge of something. PTs appeal to teachers who engage their students in the productive (speaking and writing) aspects of language rather than the receptive aspects (listening and reading). Examples of performance tests include writing, roleplays, and giving a presentation. The purpose of a PT is to evaluate a candidate on a set of *criteria*. According to McNamara (1997), a PT uses criteria, something the teacher wants and expects the candidate to do in real life, and the test is a simulated performance providing a sample of language. The function of the PT is to supply data to allow an inference about what the ITA will do later. In a presentation, the test candidate picks a topic that is similar to one he or she might teach, for example a key term that can be defined, illustrated, and explained. Teachers rate the performance in real time on a set of criteria according to some scale, in this case, from one to five. Ratings are typically done by two raters in a classroom and must be completed during the presentation, lasting from five to eight minutes.

Motivation for the Current Study

The purpose of this test validation study was to investigate to what extent each criterion on Performance Test version 8.3 (see Appendix) contributed to the purpose of the test,

which is to identify students who can be approved to teach. Criteria identified as not functioning can be eliminated or revised. Our study will investigate the following evaluation question (EQ): Are all the ten criteria of Performance Test v8.3 functioning to distinguish between Masters and Non-masters?

METHOD

Participants

Participants in this study were 146 ITA candidates who completed the 2010 summer ITA workshop. Of these, 80 were males and 66 were females. They came from 39 countries mainly China, India, South Korea, Sri Lanka, and Thailand. The majority of the ITA candidates were enrolled in doctoral programs in 29 departments mostly (about 70%) in Biology, Chemistry, Foreign Languages, Math, and Petroleum Engineering. As a result of the workshop, there were 67 masters and 79 non-masters.

Materials

ITA Performance Test v8.3 was created by Gorsuch, Meyers, Pickering & Griffee (2010). Version one, titled *The ITA Test*, was initially published in *Communicate* (Smith, Meyers, & Burkhalter, 1992). For versions two through seven, the name of the test was changed to the *ITA Presentation Test* to reflect the primary use of the test to evaluate a class presentation required for each ITA candidate. A history of the development of The ITA test versions one to six is available from Gorsuch (2006).

Version seven was based on the communicative competence theory found in Bachman & Palmer (1996); however, the test still utilized a curriculum taken from the *Communicate* textbook. In version eight, the test ceased reflecting the assumptions of *Communicate*, and attempted to more accurately reflect the curriculum exemplified in Gorsuch, Meyers, Pickering and Griffee (2010). Version eight can be seen as a shift in emphasis rather than a complete break from the past in that version eight is a shift from *a real-life approach* to an *interactional/ability approach* in which the test reflects communication as a theory rather than a perception of individual abilities (Bachman, 1990).

ITA Performance Test v8.3 (Appendix) contains 10 criteria that raters use to assess the English abilities of ITA candidates. Each criterion is titled with a brief definition of the variable given underneath the title. A 5-point Likert scale is placed under the definition of the variable for raters to assign a score. A star (*) is placed on a score of 4 for each criterion with a description of the abilities displayed by an ITA candidate at that level. Because a passing score for the performance test is at least nine 4s with only one 3, defining a score of 4 is necessary for face validity and rater reliability.

Raters.

In the spring semester prior to the summer workshop, people are recruited to work as raters and instructors at the ITA workshop. The hierarchy of recruiting raters begins with current Teaching Assistants (TAs) of the regular academic year ESL class. Because the volume of ITAs during the summer session is significantly more than the academic year, additional raters, other than the ESL TAs, are needed. Next on the hierarchy of recruitment for the workshop are raters of previous ITA workshops. Because the workshop needs instructors and teaching assistants, previous assistants are the next to be recruited. Finally, current Applied Linguistics Masters candidates are recruited to fill out

the remaining rater and assistant positions. All raters and assistants in the workshop are either Applied Linguistics Masters candidates or degree holders. Two days prior to the start of the ITA workshop, all raters and assistants are trained on how to rate the ITA Performance Test, regardless of previous workshop experience. Previous ITA performance videos are then shown to raters and assistants in order to align scores. Over the past three years, the ratio of experienced raters to non-experienced raters has been 4:2.

Procedure

Candidate scores were coded as Masters and Non-masters (Brown, 2005). For each candidate who took the Performance Test, two raters scored the candidate on each of the criteria for the Performance Test (Appendix). Each criterion is awarded a score of one to five. The two sores for each candidate on all criteria were entered into an SPSS statistical program to analyzed using a MANOVA.

Analysis

Although Schaefer (2008) and Kondo-Brown (2002) utilized the FACETS program to determine rater bias in their studies, our research suggests the use of a MANOVA for achieving similar results through analysis of the items. Analyzing the items rather than raters is better suited to a program that does not have a regular in-house staff of raters. With the ITA workshop and semester course (ESL5310), new raters are recruited yearly. A MANOVA analysis is able to analyze data from a variety of raters and assess the items on the test, not the abilities of the raters, which is more appropriate to a university level program with a changing staff of raters.

RESULTS

MANOVA analysis results for the interaction between Performance Test items and Master and Nonmaster groups can be seen in Table 1.

Table 1.

MANOVA Analysis of Interaction between Items and Groups for Performance Test 8.3.

Criteria	F-value	*p*-value
One	32.51	.00
Two	27.38	.00
Three	21.91	.00
Four	02.69	.10
Five	05.61	.02
Six	09.29	.00
Seven	09.07	.00
Eight	09.42	.00
Nine	02.76	.10
Ten	08.80	.00

Because each item on the Performance Test is expected to separate Master from Non-master candidates, non-significant results are the focus of this study. Criteria four and nine on ITA Performance Test v8.3 are above the cut point of 0.05 and are judged not significantly different and therefore are judged as not functioning as intended.

Why criteria four and nine did not function adequately

A possible explanation for the inability of criterion four to discriminate between master and non-master groups is that criterion four is a loaded item, meaning that there are multiple elements that, although related, require the rater to think about each one, which is time consuming and possibly distracting. Specifically, criterion four asks the rater to estimate cohesion in terms of grammatical structures, word choice, and transitional phrases. If a test candidate were perceived as fulfilling one element of the criterion but not the other, the rater is left with the problem of what score to assign. It is likely that raters, pressed for time and required to make a decision, tended to give a passable score just to satisfy the requirements of the test. This would result in scores that would not differentiate master from non-master.

Criterion nine may not have functioned because it was not be perceived by raters as pertinent to assessing English abilities based on the communicative theory (Bachman & Palmer, 1996). Criterion nine addresses whether candidates used a visual and whether it was used "effectively." The description associated with the item states, "Visuals can be clearly seen, candidates talk about them, explains why they are using them." The description for what deserves a score of four out of five only addresses the conscious awareness of the visual by the candidate and audience, but not the linguistic skills when presenting it.

Solutions and revisions

One possible solution is to delete criteria four and nine, leaving the performance test with eight criteria. The advantage of this solution is that an eight-criterion performance test would be easier to grade in the real-time context the test operates. The disadvantage, however, would be that criterion four contains aspects of the Communicative Language Theory that stakeholders and raters agree adds face validity to the performance test. The limitation of criterion four stated by both groups is that there are several variables that are independent of each other.

A second possible solution is to cut criterion nine because is does not require any linguistic ability to answer, and revise criterion four by dividing it into two parts, and to use each of those parts as new criterion. Because raters perceive criterion nine as a judgment of aesthetics rather than a judgment of linguistic ability, there is no challenge to delete the criterion. Dividing criterion four would ensure that important components of the curriculum presently assessed in criteria four such as the *use of grammatical structures* and *transitional phrases* to strengthen cohesion will continue to be assessed. After consultation with co-authors of the textbook in which the test appears, this was the course of action we took. There are no changes to criteria one, two, three, or ten either in content or in location in the test. Criterion four was divided into new number four and new number five and criterion nine was cut.

CONCLUSION

This test instrument validation study was conducted on a performance test that has a long history including multiple revisions. Notwithstanding, this empirical investigation found that two of the ten criteria functioned in a sub-optimal way, a matter of concern in a high stakes program. We conclude that in test instrument development and validation, there is no substitute for empirical verification. The take-home lesson is that we cannot assume that our tests are functioning just because they have a long history. Test validation, the systematic investigation of test performance in terms of test use, is a necessary exercise.

ABOUT THE AUTHORS

Dale T. Griffee Ed. D. directs the ITA training program at Texas Tech University. He teaches course evaluation, testing, and research methods in the applied linguistics MA program and ESL academic writing. Email: dale.griffee@ttu.edu Mailing address: Dale T. Griffee, Texas Tech University, Box 42071, Lubbock, Texas 79409. Phone number: (806) 742-3145

Jeremy Gevara is a Master's Candidate with the Applied Linguistics program at Texas Tech University. His research and career interests are in second language assessment and second language program evaluation. Email: jeremy.gevara@ttu.edu Mailing address: Jeremy Gevara, Texas Tech University, Box 42071, Lubbock, Texas 79409. Phone number: (806) 742-3145

REFERENCES

Alderson, J. C. (1986). The nature of the beast. In A. Wangsotorn, A. Maurice, K. Prapphal & B. Kenny (Eds.), *Trends in language programme evaluation* (pp. 5-24). Bangkok: Chulalongkorn University.

Allen, M. J. (2004). *Assessing academic programs in higher education*. San Francisco: Jossey-Bass.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bailey, K. M., & Brown, J. D. (1996). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236-256). Philadelphia, PA: Multilingual Matters.

Brown, J. D. (1995). Language program evaluations: Decisions, problems and solutions. *Annual Review of Applied Linguistics, 15*, 227-248.

Brown, J. D. (2005). *Testing in language programs*. New York: Prentice McGraw Hill.

Gorsuch, G. (2006). Classic challenges in International Teaching Assistant assessment. In D. Kaufman & B. Brownworth (Eds.), *Professional development of international teaching assistants* (p. 69-80). Washington, DC: Teachers of English to Speakers of Other Languages.

Gorsuch, G. J., Meyers, C., Pickering, L., & Griffee, D. T. (2010). *English communication for international teaching assistants*. Long Grove, IL: Waveland Press.

Graves, K. (1996). A framework of course development processes. In K. Graves (Ed.), *Teachers as course developers* (pp. 12-38). Cambridge: Cambridge University Press.

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), 17-29. Retrieved from http://studentorgs.utexas.edu/flesa/TPFLE_New/Index.htm.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Lynch, B. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University Press.

McNamara, T. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education*. Boston: Kluwer.

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13*(3), p. 241-256.

Palmer, A. (1992). Issues in evaluating input-based language teaching programs. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 141-166). Cambridge: Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Sheridan, J. D. (1991). A proactive approach to graduate teaching assistants in the research university: One graduate dean's perspective. In J. Nyquist, R. D. Abbott, D. H. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 24-28). Dubuque, IA: Kendall/Hunt.

Smith, J., Meyers, G., & Burkhalter, A. (1992). *Communicate: Strategies for international teaching assistants*. Englewood Cliffs, NJ: Prentice Hall.

Stufflebeam, D. L., & Webster, W. (1983). An analysis of alternative approaches to evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluations* (pp 23-44). Boston: MA: Kluwer-Nijhoff.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.

Vogt, W. P. (2005). *Dictionary of statistics & methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Wilkening, L. L. (1991). Teaching assistants: Training for the professoriate. In J. Nyquist, R. D. Abbott, D. H. Wulff, & J. Sprague (Eds.), *Preparing the professoriate of tomorrow to teach* (pp. 12-16). Dubuque, IA: Kendall/Hunt.

APPENDIX

ITA Performance Test v8.3

Grammatical competence

1. The ITA candidate pronounces sounds clearly enough at the word level that the listener can understand what word is intended.

1	2	3	4	5	Occasional difficulty, but usually understandable.
Low			*	High	

2. ITA uses word stress (*expectation, similar*) and does not add or drop syllables.

1	2	3	4	5	Multisyllabic words usually understandable.
Low			*	High	

3. ITA candidates uses thought groups effectively.

1	2	3	4	5	Generally listeners not aware of whether thought groups used.
Low			*	High	

Textual competence

4. ITA uses grammatical structures, word choice, and transitional phrases effectively to provide cohesion to the content (*Let me give you an example of this theory*).

1	2	3	4	5	Listener can generally follow the logic of the talk.
Low			*	High	

5. ITA gives clear definitions and examples based on audience awareness.

1	2	3	4	5	Candidate frequently inserts definitions and examples.
Low			*	High	

Sociolinguistic competence

6. ITA uses prominence.

1	2	3	4	5	Listeners are aware of important words.
Low			*	High	

7. ITA aware of listener non-comprehension by techniques such as eye-contact, wait time, and checking for comprehension. (*Does everybody understand so far?*)

1	2	3	4	5	Does at least two of the above.
Low			*	High	

8. ITA varies tone choice so as to produce a variety of rising and falling tones; not a monotone.

1	2	3	4	5	Not all rising tones, not all falling tones.
Low			*	High	

9. ITA candidate uses visuals or multimedia effectively.

1	2	3	4	5	Visuals can be clearly seen, candidate talks about them, explains why they are using them.
Low			*	High	

Functional competence

10. Candidate expands beyond audience questions by acknowledging the question, confirming understanding by repeating or paraphrasing the question, answering the question, and checking back to confirm question has been answered.

1	2	3	4	5	Candidate accomplishes at least 3 of the 4 techniques.
Low			*	High	

Recommendations for the future that the candidate can work on.