The Use of Proficiency Rubrics in an IEP Placement Test

Deirdre J. Derrick

Northern Arizona University

Abstract

Students who take language classes need to be placed into the appropriate level so that they may best benefit from instruction. In Intensive English Programs (IEPs) placement may be done in multiple ways, using prior grades, scores on a proficiency exam such as the TOEFL, or with an in-house placement test. There is limited research on how well these different methods function for placement purposes, and much of it has been done in a foreign language context (e.g., LaBlanc & Lally, 1997). The present study looks at one IEP which administers an in-house placement test, but which uses TOEFL speaking rubrics to score two of the three speaking tasks. The study uses data from two administrations of the placement test and analyzes it using multi-faceted Rasch measurement (MFRM). The goal was to determine how well the proficiency rubrics worked in conjunction with the test tasks for placement purposes. Results suggested that the proficiency rubrics did not provide fine-grained information to place examinees into one of six levels. Results also suggested that raters, as a whole, were unable to apply the rubrics consistently. These results have implications for the IEP, in terms of rater training and scale revision. They also have implications for other programs regarding the design and scoring of placement tests.

The Use of Proficiency Rubrics in an IEP Placement Test

Background

Placement tests are used “to determine for each student the position in the instructional sequence and the mode of instruction that is most beneficial” (Miller, Linn, & Gronlund, 2009, p. 38). There are several ways that placement in to language courses have been made. These include using prior grades or test scores, including scores from a high-stakes proficiency exam (e.g., TOEFL, IELTS), scores from a commercially available placement test (e.g., the TOEFL ITP, the CaMLA EPT), or scores from an in-house placement test. Using grades, particularly those from other institutions, is problematic because it can be difficult to interpret them. Test scores have been shown to be an appropriate way to place students into language classes (LeBlanc & Lally, 1997).

In language testing literature, tests are divided as being either norm-referenced (NRT; e.g., most proficiency tests) or criterion-referenced (CRT; e.g., achievement tests). The distinction between NRTs and CRTs is important because it provides information about how test scores are to be interpreted. NRTs provide information about how well examinees do in relation to each other, and CRTs provide information about how well examinees do in relation to a curriculum. The content of each type of test is guided by these two different principles. Proficiency tests are generally NRTs, though some tests such as IELTS and TOEFL are CRTs because performance scores are tied to descriptions of language ability (Carr, 2011).

Placement tests inhabit an awkward position in relation to these types. Some authors describe them as norm-referenced (Brown, 2005), some as criterion-referenced (Carr, 2011), and some as potentially one or the other (Bailey, 1998) or even both (AERA, APA, NCMA, 1999). Placement tests can have features of both types. Like criterion-referenced tests, they are developed with a particular curriculum in mind. Like norm-referenced tests, such as a proficiency test, they assess a wider range of abilities and are used to make program-level decisions (Brown, 2005). Placement tests are more like criterion-referenced tests in that their scores provide information about what examinees are able to do and not just their order of ability relative to other examinees (Davidson & Lynch, 2002). Criterion-referenced placement tests have multiple cut scores, which as used to make placement decisions into each level (Carr, 2011).

The situation vis-à-vis placement tests and proficiency tests is somewhat more complicated in the IEP examined in this study. Proficiency tests cover a broader range of ability levels than placement tests, so it may be expected that a rubric from a proficiency test would be too broad for a placement test. If this is the case, then student scores would tend to cluster at fewer score levels, and would not be spread out among all possible levels. This would make it difficult to make placement decisions, since more fine-grained information about student abilities would not be available.

Research Questions

1. Do raters use the full scale for the independent speaking task?
2. Do raters use the full scale for the integrated speaking task?
3. How do examinee logits compare with placement decisions?

Methods

Examinees

Examinees were students at a North American IEP. They were primarily Chinese and Arabic-speaking. They were either new students taking the placement test to be placed into a level or returning students taking the placement test to advance a level. Data from 247 participants from the fall 2014 (*N* = 149) and fall 2015 (*N* = 108) placement tests were used. Duplicate examinees (examinees who took both placement tests) were removed from the data set for the second (fall 2015) administration.

Raters

Raters were M.A. or Ph.D. students, full-time IEP employees, or members of the IEP assessment team. All were employed by the IEP at the time of scoring, and all had experience teaching English as a Second Language. All raters possessed a background in ESL and/or EFL. Prior to rating, raters were given a training session in which they were familiarized with the task, the rubric, and the importance of the placement decision-making process. This was followed by a calibration session in which raters were given eight essays to score. Each examinee response was then scored by two raters, and each task had a scoring leader who monitored raters and answered questions.

Speaking scores come from eight raters for the independent task and six raters for the integrated task from the fall 2015 placement test administration as well as from nine raters for the independent task and eight raters for the integrated task from the fall 2014 placement test administration. Some raters scored more than one task and more than one administration, giving a total of 23 unique raters.

Placement Test

The speaking section of the placement test has three tasks. The first task provides examinees with a picture of a process which they are asked to describe. This task is scored using an in-house rubric. The second task provides examinees with two pictures and they are asked to compare and contrast the two. This task is scored using the TOEFL iBT independent speaking rubric. The third task provides examinees with a graph, plays them a short news report, and asks them to talk about both. This task is scored using the TOEFL iBT integrated speaking rubric. Each speaking task is scored by two raters. This study uses data from the second and third speaking task.

Speaking Rubrics

Both the independent and integrated TOEFL speaking rubrics are holistic with scores from zero to four. Each score band includes a general description as well as characteristics of delivery, language use, and topic development that would describe a response at that score.

Analysis

Data analysis for the first two research questions made use of multi-faceted Rasch analysis using the FACETS computer program. The facets analyzed included examinee, rater, and task. The partial credit model was used because each task used a different rubric.

The third question was answered by performing a Spearman Rho correlation between examinee logit and level placement.

Results

Examinees showed a wide spread from 9.07 to -9.49 logits. Most examinees were between 2 and -2 logits. The mean ability of examinees was -.33 and the mean infit was .94. Examinees’ ability was slightly low for this test, but not by much, and some examinees scored either extremely high or extremely low. No examinees showed misfit, as indicated by infit values of greater than 2.

Of the two speaking tasks, the independent task was easier, with a logit of -1.55. The integrated task had a logit of 1.55. Neither essay showed misfit.

Raters ranged in severity from 2.91 to -3.72, with a mean of .00. This is a rather large spread of rater severity-leniency of 6.63 logit values. No raters showed misfit.

In both scales, the average measure and the Rasch-Andrich Thresholds increases from lower scores to higher scores, indicating that each score is measuring “more” of the target quality (Eckes, 2011).

Both scales show a “central tendency effect” (Eckes, 2011). That is, scores given by raters tend to clump in the middle categories. In the independent task, most of the raters assigned either a score of 2 (229; 47%) or a score of 3 (179; 37%). Out of 492 possible scores, 408 or 81% were either a 2 or a 3. In the integrated task, most raters tended to use the middle of the scale. A score of 1 was given 195 times or 40% of the time, and a score of 2 was given 202 times, or 42% of the time. Scores of 1 or 2 were given 397 out of 492 times, or 81% of the time.

The correlation between examinee ability and placement recommendation was .621 ($p <$ .01), which is a moderately positive correlation (Hinkle, 2003).

Relevance to PIE and Second Language Learning

This section presents implications for the IEP and for the field of language assessment. For the IEP in this study, the recommendation would be to create new rubrics that provide for more fine-grained information about examinees. It may also be worthwhile to consider using speaking tasks that mirror those listed on the Student Learning Outcomes (SLOs), namely summary and synthesis tasks. Another implication arising from this study, though not related to the primary purpose of assessing scale use, is the need for more rater training. Raters in the study have very different roles in the IEP (graduate assistant, administrator, instructor, and coordinator). These differences may have been the cause of the wide range of rater severity in logits, but better training may address this issue.

For the field of language assessment, the main implication arising from this study is the suggestion that rubrics created for proficiency tests may not be appropriate to use for tests with other purposes. This was the case in the present study, where proficiency rubrics did not provide fine-grained information about examinees. This may also be the case with other types of tests, for example achievement tests. A potential mismatch could occur if the student learning outcomes on the achievement test are not the same as those measured on the proficiency rubric.

References

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

Bailey, K. M. (1998). Learning about language assessment: Dilemmas, decisions, and directions. Boston, MA: Heinle and Heinle.

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

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford, UK: Oxford University Press.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven, CN: Yale University Press.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. New York: Peter Lang.

Hinkle, D. E., Wiersm, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.). Boston, MA: Houghton Mifflin.

LaBlanc, L. & Lally, C. G. (1997). Foreign language placement in postsecondary institutions: Addressing the problem. Proceedings from Dimension '97: *Southern Conference on Language Teaching.* Valdosta, Georgia: Valdosta State University.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.