Differential Item Functioning and Item Bias: Critical Considerations in Test Fairness

Michael Perrone
Teachers College, Columbia University

Overview of Test Fairness and Item Bias

In recent years, policy makers, administrators, and test developers in the field of second language assessment have paid considerable attention to the issue of test fairness. A fair test is one that is comparably valid for all groups and individuals and that affords all examinees an equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the test's purpose (Roever, 2005). Various aspects of fairness in testing have been highlighted in the literature, including fairness in regards to standardization, test consequences/score use, and item bias (Kunnan, 2000; Shohamy, 2000). This commentary will focus, in part, upon the issue of item bias.

Item bias has considerable ramifications at a policy, administrative, and classroom level. As such, bias can lead to systematic errors that distort the inferences made in the classification and selection of students (Zumbo, 1999). Learners who have similar knowledge of the material on a test (based on total examination results) should perform similarly on individual examination items, regardless of gender, culture, ethnicity, or race (Subkoviak, Mack, Ironson, & Craig, 1984). An examination item is considered biased if it functions differently for a specified subgroup of test-takers; in such a case, students who are equally able do not have an equal chance of success (Zumbo, 1999). A biased item measures attributes irrelevant to the tested construct (Williams, 1997). Frequently, examination items are considered biased because they contain sources of difficulty that are not relevant to the construct being measured and these extraneous sources impact test-takers' performance (Zumbo, 1999). An item might also be considered biased if it contains language or content that is differentially difficult for different subgroups of test-takers. In addition, an item might demonstrate item structure and format bias if there are ambiguities or inadequacies in the item stem, test instructions, or distractors (Hambleton & Rodgers, 1995).

Previously, a variety of methods had been proposed for detecting item biasness, including but not limited to the following: the transformed item difficulty method, the chi-square method, and the three-parameter item characteristic curve (Subkoviak et al., 1984). However, over the past decade, Differential Item Functioning (DIF), developed by the Educational Testing Service (ETS) in 1986, has become the standard of psychometric bias analysis and, as such, will be the focus of this commentary (Roever, 2005).

Differential Item Functioning

Logically, the first step in detecting test bias is to locate examination items on which one group of test-takers performs significantly better than another group (Roever, 2005). DIF is a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge of different groups of examinees (e.g., male vs. female or

Caucasian vs. African-American; Schumacker, 2005). As such, DIF aids in the identification of test items that are potentially biased. In assessing response patterns, the comparison groups (e.g., males vs. females) are initially matched on the underlying construct of interest (e.g., verbal ability or mathematics achievement). By matching groups on the measured variable, researchers/test developers are better able to determine whether item responses are equally valid for distinct groups of test-takers (Zumbo, 1999).

DIF methods therefore assess the test-takers' response patterns to specific test items. DIF occurs when a statistically significant difference is evident in the probability that test-takers from the two distinct groups (e.g., males and females), who have the same underlying ability on the measured construct, demonstrate differing probabilities of correctly answering the item (Zumbo, 1999). As stated, examinees' ability levels are based upon their total scores on the examination. As such, the DIF analysis of one specific test item is as independent as possible from the DIF analyses of the other test items (Zumbo, 1999)

To reiterate, a test item is considered to be biased when a dimension on the examination is deemed to be irrelevant to the construct that is being measured, placing one group of examinees at a disadvantage in taking the examination (Hambleton & Rodgers, 1995). Thus, if DIF is *not* evident for an item, then there is no item bias. Conversely, DIF is required but is not sufficient for item bias. That is, if DIF is apparent, then its presence is not sufficient to declare item bias. An item might show DIF, but not be considered biased if the difference is a result of the actual difference in the groups' ability to respond to the item (i.e., if one group of test-takers is at a high level and the other group of test-takers is at a low level, the lower group would perform significantly lower; Roever, 2005). If test-takers differed in knowledge, a difference in item responses would be expected. Consequently, a difference in the performance of groups of examinees with different abilities on specific items is not indicative of test bias, but rather of item impact (Schumacher, 2005).

Upon seeing evidence for the occurrence of DIF, one would need to apply subsequent item-bias analyses (e.g., empirical evaluation or content analysis) in order to determine if item biasness is present (Zumbo, 1999). Only when differences in a group's ability to respond to a test item are caused by construct-irrelevant factors can DIF be considered as bias. In items exhibiting test bias, an additional construct is evident, apart from the construct that the items are supposed to measure (Roever, 2005).

Summary

The topics of test and item bias and DIF have critical political, social, and ethical implications for L2 test administrators, developers, policy makers, and examinees. Even though these topics have been the focus of much discussion and debate on a political level, there remains a relative lack of well-constructed, empirical research in the field of language testing. The study of item bias and DIF is critical, as such research helps provide an empirical foundation for the identification and subsequent elimination of exam items that appear to be relatively more difficult for one group of test-takers than another (Zumbo, 1999). Further research on these issues will allow us to comprehend more fully the possible substantive interpretations that can be made by focusing upon test items considered to be biased. In addition, subsequent research can help us understand in greater depth the factors that contribute to DIF (i.e., which examinee background variables interact with which test items in which way?). Through such a lens, test developers will be better able to construct examinations that are fair and appropriate measures of test-takers' knowledge of the examination material.

REFERENCES

Hambleton, R., & Rodgers, J. (1995). Item bias review. *Practical Assessment, Research, and Evaluation*, *4*(6). Retrieved November 18, 2006, from http://PAREonline.net/getvn.asp?v=4&n=6

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge, UK: Cambridge University Press.

Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing*. Retrieved November 18, 2006, from the University of Hawai'i System Web site: http://www2.hawaii.edu/~roever/brownbag.pdf

Schumacker, R. (2005). *Test bias and differential item functioning*. Retrieved November 18, 2006, from http://www.appliedmeasurementassociates.com/White%20Papers/TEST%20BIAS%20AND%20DIFFERENTIAL%20ITEM%20FUNCTIONING.pdf

Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 15-19). Cambridge, UK: Cambridge University Press.

Subvokiak, M., Mack, J., Ironson, G., & Craig, R. Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21*, 49-58.

Williams, V. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10*, 253-267.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-like (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation.

Michael Perrone is a doctoral student in Applied Linguistics. His research interests include L2 washback and L2 motivation.