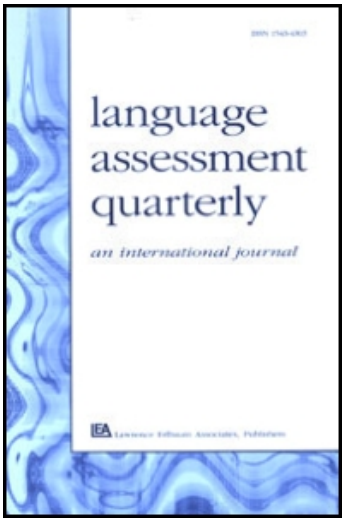

Language Assessment Quarterly
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653669

Test Fairness, Test Bias, and DIF

Antony John Kunnan

To cite this Article Kunnan, Antony John(2007) 'Test Fairness, Test Bias, and DIF', Language Assessment Quarterly, 4: 2, 109 — 112
To link to this Article: DOI: 10.1080/15434300701375865
URL: http://dx.doi.org/10.1080/15434300701375865

LANGUAGE ASSESSMENT QUARTERLY, *4*(*2*), 109–112

EDITORIAL

Test Fairness, Test Bias, and DIF

I became personally interested in test bias studies in the late 1980s as a doctoral student at UCLA following the publication of the Chen and Henning (1985) study on test bias, arguably the first such study in the field of language assessment. Miyuki Sasaki and I subsequently replicated the study, which we jointly presented at the Second Language Research Forum Conference held at UCLA in 1989 and then developed separate publications (Kunnan, 1990, 1992; Sasaki, 1991). I then became interested in the concept of test fairness and argued that it should be connected to validity at the Language Testing Research Colloquium in Tampere, Finland, in 1996 (Kunnan, 1997). I proposed that test fairness has to be related to test validity and test validation. I later developed a *test fairness framework* (Kunnan, 2000, 2004) in which I presented *absence of bias* as a test quality and argued that one way of reducing or eliminating bias would be through studies that examined test items for differential item functioning (DIF).

TEST FAIRNESS

Test Fairness Framework

Figure 1 presents the Absence of Bias quality from the Test Fairness Framework (TFF; Kunnan, 2004). Although test fairness is a central quality, absence of bias is shown as a contributing and interrelated quality (along with other interrelated test qualities). I am ignoring the other qualities in this discussion because I want to focus on absence of bias. I argue that absence of bias investigations can be used as part of the TFF in evaluating tests. Three interrelated aspects contribute to absence of bias quality:

1. *Content or language variety*: This type of bias refers to content or language or dialect that is offensive or biased to test takers from different backgrounds. Examples include content or language stereotypes of group members and overt or implied slurs or insults (based on gender, race and ethnicity, religion, age, native language, national origin, and sexual orientation) or choice of dialect or variety that is biased to test takers.
2. *Group performance*: This type of bias refers to difference in performances and resulting outcomes by test takers from different group memberships. Group differences could occur among salient groups (e.g., gender, race and ethnicity, religion, age, native language, national origin, and sexual orientation) on test tasks and subtests.
3. *Standard setting*: This type of bias refers to standard setting in terms of the criterion measure and selection decisions and how these decisions affect different test taking groups.

FIGURE 1 Absence of Bias from the Test Fairness Framework.

AERA, APA, NCME (1999) Standards

In the recent *Standards* (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/NCME], 1999), in the chapter titled "Fairness in Testing and Test Use," the authors state by way of background that the

> concern for fairness in testing is pervasive, and the treatment accorded the topic here cannot do justice to the complex issues involved. A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. (p. 73)
>
> The first two characterizations . . . relate fairness to *absence of bias* and to *equitable treatment of all examinees* in the testing process. There is broad consensus that tests should be free from bias . . . and that all examinees should be treated fairly in the testing process itself (e.g., afforded the same or comparable procedures in testing, test scoring, and use of scores). The third characterization of test fairness addresses the *equality of testing outcomes* for examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature. A more widely accepted view would hold that examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership. . . . The fourth definition of fairness relates to *equity in opportunity to learn* the material covered in an achievement test. There would be general agreement that adequate opportunity to learn is clearly relevant to some uses and interpretations of achievement tests and dearly irrelevant to others, although

> disagreement might arise as to the relevance of opportunity to learn to test fairness in some specific situations [italics added]. (AERA/APA/NCME, 1999, p. 74)

The *Standards* document presents 12 standards for fairness. The standards that are relevant are summarized here:

- Validity evidence collected for the whole test group should also be collected for relevant subgroups
- A test should be used only for the subgroups for which evidence indicates that valid inferences can be drawn from test scores
- When DIF exists across test taker characteristic groups, test developers should conduct appropriate studies
- Test developers should strive to identify and eliminate language and content that are offensive by subgroups except when necessary for adequate representation of the domain
- When differential prediction of a criterion for members of different subgroups are conducted, regression equations (or appropriate equivalent) should be computed separately for each group
- When test results are from high-stakes testing, evidence from mean score differences between relevant subgroups should be examined and if such differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevance variance.

These frameworks, definitions, and standards suggest that fairness studies need to be mandatory so that tests can be free of bias and ultimately become fair tests. Personally, it was only through the development of the *Test Fairness Framework* and the development of the *Standards* (AERA/APA/NCME, 1999) that I began to see how DIF studies can contributed to the larger view of test fairness instead of DIF studies being seen as one-off studies. It is this view that I am proposing for this special issue as well: to view DIF detection methods as a way of eliminating or reducing bias keeping in mind the ultimate goals of absence of bias and of test fairness.

THIS SPECIAL ISSUE

This special issue, devoted to DIF detection methods in language assessment, is arguably the first special issue in a language assessment journal devoted to this methodology. The first article, by Tracy Ferne and André Rupp, reviews research on DIF in language testing conducted primarily between 1990 and 2005 with an eye toward providing methodological guidelines for developing, conducting, and

disseminating research in this area. The article contains a synthesis of 27 studies, and it presents and discusses the features of the DIF detection methods that have been applied, the reporting of DIF effects, and the explanations for and consequences drawn from DIF results.

Gary Ockey's article reports on a study regarding English language learners' test performance on a test in a subject matter area such as mathematics. The article considers whether construct irrelevant variance in a math test could result from English, the language in which the test is presented. Carsten Roever's study investigates DIF in a 36-item test of English as a Second Language pragmalinguistics assessing language learners' knowledge of implicature, routines, and speech acts. Ardeshir Geranpayeh and Antony Kunnan's contribution reports on study that examines the Certificate in Advanced English Examination, a test developed by Cambridge ESOL Examinations, for DIF in terms of age. Bruno Zumbo's commentary rounds out the special issue with a discussion of the three generations of DIF analyses: where it "has been, where it is now, and where it is going."

We have tried to present the issues related to DIF in a conceptual and less technical manner as possible, keeping in mind that all of our readers are not measurement experts. We hope we have succeeded in convincing you that research using DIF detection methods is a productive and useful way to ensure test fairness.

Antony John Kunnan
June 2007

REFERENCES

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

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*, 155–163.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741–746.

Kunnan, A. J. (1992). Response to "The limits to biased item analysis". *TESOL Quarterly, 26*, 598–602.

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 85–105). Jyväskylä, Finland: University of Jyväskylä.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp.1–13). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages Conference Papers, Barcelona, Spain* (pp. 27–48). Cambridge, UK: Cambridge University Press.

Sasaki, M. (1991). A comparison of two methods for detecting DIF in an ESL placement test. *Language Testing, 8*, 95–111.