The Author Responds . . .

ANTONY JOHN KUNNAN
The University of Michigan

Though Stuart Luppescu seems to share the common concern of developing and using tests and test items without differential item functioning (DIF), there are serious problems with his evidence, arguments, and interpretation regarding my study, in particular, and DIF research in general.

At the very outset in his commentary, in the first paragraph, he concludes with a sweeping statement: "the *fact* that statistical and subjective procedures for identifying DIF cannot be relied upon dilutes the significance of his conclusions and of the conclusions of *nearly all* [italics added] studies which investigate the causes of DIF." Before responding substantially to this statement, I would first like to know what research evidence he has to support the assertion that there is a "fact" about statistical and subjective procedures for identifying DIF which compels him not to rely upon them. Second, I would like to know which studies are excluded from his assertion since he leaves the door open with "nearly all" studies.

Luppescu pointedly criticizes the method I used in my study. My study was a posteriori analysis of an ESL placement test with 150 multiple-choice items. I used a method similar to the Delta-plot method within an overall one-parameter item response or Rasch model (see Angoff, 1982; Angoff & Ford, 1973; Chen & Henning, 1985, for procedural details). In addition to the justification I provided in my article for this approach, I would add that the Rasch model is quite suitable for multiple-choice items (see Henning, 1989). Luppescu specifically argues that because I used the 95% confidence interval for the regression plot to identify DIF based on item-difficulty indices, "we expect 5% of the items, even if there is no bias, to be selected by this method." This is misinterpreting and confounding the level of significance chosen for statistical tests of hypotheses with the percentage of items identified as having DIF. The two are not related and comparable and, therefore, cannot be used in the manner Luppescu does. Wainer's (1991) "isthmus of acceptance" proposal, however, is an interesting one in this regard.

Luppescu further argues that his simulation study of 1000 subjects and 75 items with' no bias identified 6 items as containing DIF. He then concludes that since there was no generated bias in his data, "ordinary, expected, stochastic variation" identified DIF for items

that did not have such a characteristic. Based on this finding, he argues that in my study, of the 36 items that exhibited DIF, "at least, 7 or 8 items do not display DIF but were selected by the procedure because of ordinary variation." Once again, there are several problems with this conclusion. First, Burrill's (1982) excellent review of simulation studies shows that identification of items that have induced DIF is quite accurate (see, for example, McCauley& Mendonza, 1985; Rudner, Getson & Knight, 1980a, 1980b; Subkoviak, Mack, Ironson, & Craig, 1984). These studies show that studies based on simulated data do not have an inherent problem in the way Luppescu argues. Besides, as Burrill (1982) correctly points out, simulation studies have their limitations.

Second, Luppescu incorrectly assumes that if a statistical procedure identifies items that display DIF, then those items are biased. He argues, referring to my study, that "some items identified by statistical procedures as biased are actually not, and that it is impossible to tell by nonstatistical procedures which are and which are not." The argument he makes here implies that he sees statistical procedures and nonstatistical procedures as two separate ways of identifying DIF. Again, this is a misreading of the ways in which the two procedures work: Statistical procedures are empirical, internal methods that strictly examine items for DIF in "context" (that is, item sets must be homogeneous, belonging to the same content or construct), whereas judgments (by experts or test reviewers) are external and often made at the item level, ignoring context or construct (see Coffman, 1982, for directions to review panels judging the Iowa Tests of Basic Skills). Thus, these two approaches could yield different though valuable results, but reliance on any one of the approaches would be wrongheaded.

In addition, the measurement literature is full of caution regarding total dependence on statistical as well as nonstatistical approaches in identifying DIF. As Shepard (1982) states, "there is no foolproof statistical bias detection method. Item bias techniques themselves require validation" (p. 22). This point of view has been articulated through validation and reliability studies (for example, Hoover & Kolen, 1984; and Shepard, Camilli, & Williams, 1985) and through recent attempts to find the most appropriate method (for example, Ryan, 1991; Swaminathan & Rogers, 1990; and Wainer, Sireci & Thissen, 1991). The use of judgments, too, has been questioned (see Reynolds, 1982, and Sandoval & Miille, 1980) though test publishers use item-review forms and test-sensitivity reviews (see Berk, 1982), sometimes to the exclusion of statistical procedures. Therefore, for the best results, both statistical and judgmental approaches should be used in combination.

This combinatory approach was used in my study: First, items that were aberrant for the different native language and gender groups were identified through statistical procedures using the Rasch model. These items were then examined for construct or content differences from other items in the set so that potential sources of DIF could be hypothesized through nonstatistical procedures. Three potential sources of DIF (instructional background, major field, and native language) for 22 (61%) out of the 36 items were identified, leaving 14 (39%) of the items with no hypotheses and explanation. Thus, my study, in general, and my conclusions, in particular, were exploratory and speculative (similar in approach to Scheuneman, 1987) rather than confirmatory.

To conclude, a DIF study would be seen as critical to language testing research when it is conceptualized as a special case of construct validation because if a test or test items exhibit DIF, not only do the test or the test items have the potential for bias but test invalidity could occur. And, this would mean that test scores would be distorted for all groups, and decisions made on the basis of such results could be invalid. This conceptualization could help transform the role of DIF research from the narrow focus of being fair to all groups to the broad view of validating tests and test-score use. Thus, DIF research could contribute to construct validation as much as other validation studies which model test performance, test methods, and test-taker characteristics (for example, Kunnan, 1992). Then, researchers like Luppescu would consider DIF research not merely as a way of developing "culture-fair" tests without "biased" items but as a way of developing acceptable construct validity for test-score use.

■

REFERENCES

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: The Johns Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95-106.

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Burrill, L. E. (1982). Comparative studies of item biased items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 161-179). Baltimore: The Johns Hopkins University Press.

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

Coffman, W. E. (1982). Methods used by test publishers to "debias" standardized tests: Riverside Publishing Company / Houghton Mifflin. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. *240-*255). Baltimore: The Johns Hopkins University Press.

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26,* 91-97.

Hoover, H. D., & Kolen, M. J. (1984). The reliability of six item bias indices. *Applied Psychological Measurement, 8,* 173-181.

Kunnan, A. J. (1992, April). *A case for construct validation of tests through structural modeling.* Paper presented at the Department of Linguistics Colloquium Series, University of Michigan, Ann Arbor.

McCauley, C. D., & Mendonza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9,* 389-400.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199-227). Baltimore: The Johns Hopkins University press.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213-233.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1-10.

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal* of *Educational Measurement, 28,* 325-337.

Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48,* 249-253.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24,* 97-118.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore: The Johns Hopkins University Press.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximating techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21,* 49-58.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Wainer, H. (1991). The isthmus of acceptance: A graphical tool for function-based item analysis and test construction. *Journal of Educational Statistics, 16,* 109-124.

Wainer, H., Sireci, S. S., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Research Issues

The *TESOL Quarterly* publishes brief comments on aspects of qualitative and quantitative research. For this issue, we asked three researchers to address the following question: How should qualitative researchers in our field understand reliability and validity?

Edited by GRAHAM CROOKES
University of Hawaii at Manoa

Validity and Reliability in Qualitative Research on Second Language Acquisition and Teaching

Two Researchers Comment. . .

DONNA M. JOHNSON and MURIEL SAVILLE-TROIKE
University of Arizona

Because research on second language acquisition and teaching (SLAT) draws on and contributes to a variety of disciplines, it is important to study differing views on the nature of inquiry. Researchers and teacher-researchers in this field should be able to read, assess, conduct, and benefit from research with an understanding of different views about what constitutes high-quality inquiry. Eisner and Peshkin (1990) suggest that being bimethodological or multimethodological is a mark of scholarly sophistication. This idea is worth considering for SLAT students who need to know about methods and standards of inquiry in linguistics, education, the humanities, anthropology, psychology, sociology, and so on. We focus here on the notions of validity and reliability as standards in research. Although our own research perspective is essentially qualitative in nature, we will argue for the potential utility of auxiliary quantitative procedures in achieving these standards. We will also argue that qualitative procedures are important for establishing the validity of research conducted from an essentially quantitative perspective. In other words, the two approaches should be seen as complementary rather than mutually exclusive (see, for example, Jaeger, 1988).

Notions of validity differ substantially in different research traditions, but the generally accepted view derives from a positivist-realist