# BRIEF REPORTS AND SUMMARIES

The *TESOL Quarterly* invites readers to submit short reports and updates on their work. These summaries may address any areas of interest to *Quarterly* readers. Authors' addresses are printed with these reports to enable interested readers to contact the authors for more details.

**Edited by GAIL WEINSTEIN-SHR**
*University of Massachusetts at Amherst/*
*Temple University*

## *DIF in Native Language and Gender Groups in an ESL Placement Test*

ANTONY JOHN KUNNAN
*University of California, Los Angeles*

■ The relative cultural fairness of educational tests has been a fertile field for researchers. The primary concern of such studies has been tests or test items that behave differently for different native-language, cultural, ethnic, and/or gender groups. These studies are popularly known as test bias studies but the term *test bias* is misleading and inaccurate. The term assumes that the researcher is examining tests or test items that are biased while what we know from test-taker responses is only that there are differences in performances by different individuals or groups that could be due to several reasons, only one of which is bias. A more accurate term, therefore, that has been used recently in testing literature, is *differential item functioning* (DIF), referring to the way items function differently for individuals or groups of test takers who have similar abilities.

Among the few studies on DIF in the field of second/foreign language testing reported recently are the following: Chen and Henning (1985) examined the Winter 1985 version of the ESL placement examination (ESLPE) at the University of California, Los Angeles (UCLA) to determine the nature, direction, and extent of bias present for members of the Spanish and the Chinese native language groups; Zeidner (1986, 1987) investigated the English Language Aptitude Test (ELAT) used routinely for student selection and placement in Israel; Spurling (1987) studied the "fair use" of the Marin Community College English admissions and placement test; Hale (1988) reported on the interaction of student major-field and text content in the reading comprehension section of the Test of English as a Foreign Language (TOEFL); and Angoff (1989) tested the hypothesis that items of the TOEFL that contain references to people, places, regions, etc., of the U.S. tend to favor test takers who have spent some time living in the United States. Kunnan and Sasaki (1989) extended the Chen and Henning (1985) study by including five native language groups in their examination of the Fall 1987 version of the ESLPE. Their study, like that of Chen and Henning (1985), identified several vocabulary and grammar items that favored certain native language groups. Other studies that have investigated performance across native language groups include Alderman and Holland (1981), and Oltman, Stricker, and Barrows (1988).

## METHOD

The present study is concerned with the identification of differential item functioning among four native language groups and the two gender groups in the Fall 1987 version of UCLA's ESLPE. (This version is different from the Winter 1985 form used in the Chen and Henning (1985) study in terms of the actual items though the composition of the test is the same.)

The sample for the study was 844 nonnative-speaking entering students at UCLA. These students were from 72 countries, with 61 language backgrounds, pursuing 76 academic specializations. The native language groups analyzed were the four largest ones: Chinese (262); Spanish (81); Korean (76); and Japanese (59), for a total of 478. In terms of gender, the sample was distributed as follows: male, 478 and female, 347 (for a total of 825; 19 test takers did not identify their sex). In its greater variety of native languages and the inclusion of gender, this study extends the work by Chen and Henning (1985), and Kunnan and Sasaki (1989).

The test instrument, the ESLPE, consists of 150 items in five 30-item subtests and one 20-minute composition. The five subtests in this test are listening comprehension, reading comprehension, grammar, vocabulary, and writing error detection. The data for the study came from the 150 multiple-choice items that were dichotomously scored for our analyses.

## RESULTS

The one-parameter Rasch model from Item Response Theory (IRT) group, which calibrates item difficulty estimates, was used for this study.[1] The analyses did not take into account the possibility of guessing. Item difficulty estimates were determined for all items using the BICAL unconditional maximum likelihood estimation procedure for the total sample. An SAS least-squares regression procedure using the item difficulty estimates was used to plot all items against all combinations of native language and gender groups. Items that were placed outside the 95% confidence interval of the regression plot were considered to be items that displayed DIF. These items were then examined in detail so that sources of DIF could be hypothesized.

[1] Similar use of Rasch analysis in identifying DIF was made earlier by Chen and Henning (1985), and Madsen and Larson (1986). Besides, the unidimensionality assumption underlying the Rasch model application appeared to be satisfied for earlier, though similar, data by Henning, Hudson, and Turner (1985) and for the present data by Davidson (1988). For more details about IRT principles, assumptions, modeling, and applications, see Henning et al. (1985), for a less technical elaboration, and Muthen and Lehman (1985) and Hambelton and Swaminathan (1985), for technical expositions.

Results of the analysis indicated the following: Thirteen items displayed DIF in the native language group analysis, and twenty-three items displayed DIF in the gender group analysis.[2]

## Native Language Group Analysis

The three items that favored the Japanese group and the three items that favored the Chinese group were all grammar items but different ones. These grammar items tested the appropriate use of grammar points, such as the definite article, a preposition, or verb tense. These items may have been easy for these two language groups because of their instructional background and/or familiarity with discrete-point grammar testing in their home countries. Thus, the source for this type of DIF can be hypothesized to be *instructional background* both in terms of test content (grammar-based teaching and testing) and test method (multiple-choice format).

The four items that favored the Spanish group were all vocabulary items. The words that were tested were *hypothetical, implication, elaborate,* and *alcoholics.* All these words have Spanish cognates making the items potentially easy for this group. This finding is similar to that of the Chen and Henning (1985) study, where the words that favored the Spanish group were *approximate, animated, maintain,* and *obstruct.* Thus, the source for this type of DIF may be hypothesized to be cognates, a test-content facet based on test-taker *native language.*

The potential sources of DIF for two grammar items, one that favored the Japanese group and one that favored the Chinese, and one vocabulary item that favored the Korean group could not be identified.

## Gender Analysis

The 20 items that favored the male group came from all sections of the test: 7 in listening, 4 in reading, 3 in grammar, 4 in vocabulary, and 2 in writing error detection. The 11 listening and reading items were based on the listening and reading passages respectively and, therefore, the content of the passages could be a source of the DIF. These items were based on passages from business, culture/anthropology, and aerospace engineering passages. These subject areas seem to favor the male group. In addition, one of the vocabulary items tested, *simulates*, may be used frequently in engineering/science classes. Out of 478 male students who took part in this study, 72% indicated that they were engineering/science majors compared with 24% of the 347 female students. The potential source of this type of DIF may have been test-taker *major field*, a test-content facet.

The potential source for DIF for 3 grammar, 3 vocabulary, and 2 writing

[2] The descriptive results showed that the test is internally consistent across all sections for all the native language groups and gender groups (KR-20: .95) and that the mean scores for the Korean and the Chinese groups and the male gender group was slightly higher than the other groups.

error detection items could not be hypothesized. In addition, the source for DIF for 3 items that favored the female group, one each from listening, vocabulary and writing error detection, could not be hypothesized.

In summary, in both analyses, out of the 150 items in the ESLPE, 36 items (24%) with no overlap between native language and gender groups were identified as displaying DIF. Potential sources of DIF for 22 of these items (61%) were hypothesized, leaving 14 items (39%) for which there were no hypotheses.

## DISCUSSION

This study identified three potential sources of DIF for both native language and gender groups: instruction, native language, and major field. But identifying potential sources of DIF is only the first step in this kind of analysis. The next step, determining what to do with items that display DIF and how to compensate test takers for such "bias" remains to be dealt with. Two procedures can be used here: (a) sources causing DIF can be examined, hypothesized and causally related, if possible, and (b) items displaying DIF can be improved or discarded. The second procedure should only be resorted to by test writers and administrators as a short-term plan: to remedy a test in use, or when there are not enough resources for a detailed study. The first procedure, which will have a more lasting effect, is recommended otherwise.

Let us use the first procedure for this study and examine the instructional background effect. This effect may be due to certain language teaching methodologies. For example, language learners who are taught through a teaching methodology that focuses on grammar and multiple-choice, discrete-point testing may find such items (as on the ESLPE) easier to answer (Farhady, 1979). This sort of practice effect can be reduced if a broad range of test content and formats is presented so as not to favor test takers from any one instructional background.

The problem of the major-field effect is easier to handle. For example, it is possible to bring about a balance in content on the basis of a test-taker background questionnaire administered to native language and gender groups that reveals reading interest, academic background, career interest, and so forth so that content that is favorable to one group does not dominate the test. Another way of balancing content would be to invite test writers from native language and gender groups who are represented in the test to balance cultural and gender-related material. (See Hale, 1988, for other suggestions.) Through these means, DIF based on content can be reduced considerably in a test.

Finally, let us examine the native language source for DIF, the cognate effect. Any language test can face this problem as specific language family ties come into play: English is more closely related to Romance and Germanic languages than, for example, Chinese or Dravidian languages. Therefore, students with a native language that is closely related to English will tend to be favored on ESL tests. Many researchers argue that this is a fact of language learning and, therefore, the question of discarding DIF items that have cognates (a linguistic effect) should not arise. Even so, many issues remain unresolved: How many questions in a 150-item test should rely on cognates? How many in each section (e.g., vocabulary, grammar)? Should the cognates reflect written language or oral language? And so forth.

## CONCLUSION

This study shows that a placement test cannot only be examined for items that display DIF by using an Item Response Theory, but also the identification of potential sources for these DIF items can be attempted and short- and long-term remedial measures to reduce DIF can then be proposed. A more extensive use of test-taker characteristics, not used in this study, will strengthen this approach. Methodologically, DIF studies could also be undertaken using the Mantel-Haenszel approach (Holland & Thayer, 1988) or Structural Equation Modeling (Muthen, 1989). Finally, a practical need for this type of study is clear: If a substantial number of items in a test display DIF, test scores could be unreliable and invalid for all groups of test takers, not just the affected groups; most tests still make placement decisions based on norm-referenced rather than criterion-referenced statistics. An example of the usefulness of this kind of study is the revision process the ESLPE has been undergoing for the past year, partly as a result of this study.

## REFERENCES

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the TOEFL* (TOEFL Research Report No. 9). Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1989). *Context bias in TOEFL* (TOEFL Research Report No. 29). Princeton, NJ: Educational Testing Service.

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

Davidson, F. (1988). *An exploratory modeling survey of the trait structures of some existing language test data sets.* Unpublished doctoral dissertation. University of California, Los Angeles.

Farhady, H. (1979). The disjunctive fallacy between discrete-point and integrative tests. *TESOL Quarterly, 13*(3), 347-357.

Hale, G. A. (1988). *The interaction of student major-field group and text content in TOEFL reading comprehension* (TOEFL Research Report No. 25). Princeton, NJ: Educational Testing Service.

Hambelton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the unidimensionality for language tests. *Language Testing,* 2(2), 141-154.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-146). Hillsdale, NJ: Lawrence Erlbaum.

Kunnan, A. J., & Sasaki, M. (1989, March). *Item bias in UCLA's ESL Placement examination*. Paper read at the Ninth Second Language Research Forum, Los Angeles.

Madsen, H. S., & Larson, J. W. (1986). Computerized Rasch analysis of item bias in ESL tests. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 47-67). Washington, DC: TESOL.

Muthen, B. (1989). Latent variable modeling in heterogenous populations. *Psychometrika, 54*, 557-585.

Muthen, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*(2), 133-142.

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). *Native language, English proficiency, and the structure of the TOEFL* (TOEFL Research Report No. 27). Princeton, NJ: Educational Testing Service.

Spurling, S. (1987). The fair use of an English language admissions test. *The Modern Language Journal, 71*(4), 410-421.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing*, 3(1), 80-98.

Zeidner, M. (1987). A comparison of ethnic, sex and age bias in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*(1), 55-71.

*Author's Address:* Department of TESL and Applied Linguistics, 3300 Rolfe Hall, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90024-1531

## *The Effect of Syntax, Speed, and Pauses on Listening Comprehension*

EILEEN K. BLAU
*University of Puerto Rico*

■ Whether or not we fully accept Krashen's theory of second language acquisition (SLA), few would deny that comprehensible input (CI), in conjunction with other factors, is an essential ingredient for SLA. Acquisition is fueled by exposure to input that is somehow rendered comprehensible either by the opportunity for negotiation of meaning via interaction or through the aid of characteristics of the input itself (Long, 1983; Snow as cited in Bohannon & Warren-Leubecker, 1985). In one-way input, i.e., lectures or a listening lab, the characteristics of the input itself are important. Long (1985) found that nonnative speakers (NNS) comprehended a foreigner-talk (FT) version of a lecture significantly better than an unmodified version. The FT version included rephrasings and restatements, was syntactically slightly less complex, and was delivered at a somewhat slower rate. However, it is not possible to determine which of the modifications (or combination of modifications) is responsible for the positive effect.

Along with rephrasings, restatements, and simplification, the notion that slowing down the flow of speech is one of the characteristics of input that