Linguistic and cultural bias in language proficiency tests

Zheng Chen and Grant Henning

Language Testing 1985; 2; 155
DOI: 10.1177/026553228500200204

Linguistic and cultural bias in language proficiency tests

Zheng Chen and **Grant Henning** *University of California, Los Angeles*

The extent to which language proficiency/placement tests may be biased for or against examinees from particular language or cultural groups has never to our knowledge been the focus of empirical research. The purpose of the present study has been to examine the English as a Second Language Placement Examination (ESLPE) employed at the University of California, to determine the nature, direction and extent of bias present for members of two linguistically and culturally diverse subgroups of the sample of examinees. By comparison of the response patterns of 34 native speakers of Spanish and 77 native speakers of Chinese from among a total sample of 312 students tested with one form of the test, it was possible to identify test items exhibiting bias in their respective skill domains. Included is a discussion of the nature, direction, extent and implications of the bias detected.

Considerable controversy has arisen around the fairness of tests used for selection decisions. Concern has centred primarily around cultural, ethnic and gender bias at the test and item levels (Peterson and Novick, 1976; Cronbach, 1976; Linn, 1976; Scheuneman, 1979; Berk, 1982; Hambleton and Swaminathan, 1985). Numerous statistical methods have been proposed for the detection and removal of bias in a variety of measurement contexts. Most of the current, successful techniques have involved applications of item response or latent trait measurement theory (Wright *et al.*, 1976; Shephard *et al.*, 1981; Linn *et al.*, 1981).

To date little research has been directed to the study of systematic bias in language tests used for decisions of university admission, placement into or exemption from remedial or preparatory language classes. Potential sources of such bias might be group differences in examinees, such as differences in native language background, type of prior educational system, or proposed specialization of study. Such background differences in groups of examinees might easily be demonstrated to differential impact testing outcomes.

A secondary concern of a more philosophical or sociopolitical nature is what to do with bias once it has been detected. Should 'biased' items be eliminated, neutralized with items said to be biased in the opposite direction, or ignored as representative of some necessary target behaviour? Certainly answers to these questions will vary in differing administrative contexts, but the questions need to be raised nonetheless.

The present study has been concerned with the identification of cultural or linguistic bias present in an English language proficiency test used for placement into or exemption from intensive English language instruction at the university level. It was felt that a test of bias for or against members of particular language groups would be more sensitive the larger and more linguistically different those groups might be found to be. Accordingly, the decision was made to focus the study on native speakers of Chinese in contrast with native speakers of Spanish. The purpose has been to identify items for which a differential, irregular probability of success was present for members of one or the other of the two groups. Use has been made of Rasch Model difficulty estimates (Wright and Stone, 1979; Rasch, 1960) in a manner analogous to the delta-plot technique proposed by Angoff (Angoff and Ford, 1973). The unidimensionality assumption underlying Rasch Model application appeared to be satisfied for the present response data set (see Henning *et al.*, 1985).

I Method

1 *Sample*

312 entering students at the University of California, Los Angeles participated in the winter 1985, administration of the English as a Second Language Placement Examination (ESLPE). These students were from more than 30 native language backgrounds, and were pursuing more than 20 academic specializations. From among the larger sample of examinees, 77 native speakers of Chinese and 34 native speakers of Spanish were identified. While included in the Chinese speaking group were speakers of both Mandarin and Cantonese dialects, it was felt that the two dialects were sufficiently similar by comparison with the Spanish language to be considered together for purposes of this study. Similarly the Spanish speakers were from several different countries with distinct dialects, but they were considered to represent one language group for purposes of the bias study.

2 *Instrumentation*

The test instrument employed in the study was the ESLPE. This exam consisted of 150 items in five, 30-item subtests, and one 20-minute composition. For purposes of the analyses described here, only the item-formatted subtests were used, and not the composition task. Overall internal consistency reliability estimated by the Kuder-Richardson Formula 20 method was 0.96. The five individual subtests exhibited reliabilities on the order of 0.83–0.88. The five multiple-choice subtests in order of administration were listening comprehension, reading comprehension, grammar accuracy, vocabulary recognition, and writing error detection. Further descriptions of the instrument are available elsewhere (Henning *et al.*, 1985; Fallis *et al.*, 1985).

3 *Procedures*

Following the identification of members of the Chinese and Spanish language groups, raw score comparisons were made of mean performance by each group on total test and each subtest. Rasch Model difficulty calibrations were determined for all test items using the BICAL unconditional maximum likelihood estimation procedure (Wright and Stone, 1979). Difficulty calibrations were derived from analysis of responses of each language group separately and then from the analysis of the combined responses of the two groups. Mean Rasch difficulty estimations were calculated for both groups on total test and each subtest. Rasch difficulty estimates were correlated for the two groups on total test and each subtest. Item difficulty estimates were plotted for the two groups and 'biased' items were identified as outliers according to a least-squares regression procedure. Biased items were tallied according to direction of bias within each subtest. Item response patterns were also compared for fit to the Rasch Model between the two language groups and across the five subtests of the test battery.

II Results

Means and standard deviations for each language group in the five subtests and test total are reported in Table 1.

From Table 1 it can be seen that the Chinese group outscored the Spanish group in every subtest except the vocabulary subtest. When the morphological similarities between English and Spanish are taken into account, some bias favouring the Spanish speakers might be expected on the vocabulary recognition portion of an ESL proficiency test.

Table 1 Means and standard deviations for two language groups on five subtests and total ESLPE

Subtest	Chinese (N = 77)		Spanish (N = 34)	
	$\bar{X}$	s	$\bar{X}$	s
Listening	19.1	4.67	15.9	4.29
Reading	23.0	5.01	20.3	5.54
Grammar	24.7	3.63	20.9	5.23
Vocabulary	19.2	5.14	19.7	5.01
Writing error detection	16.3	4.20	14.3	4.83
Total	102.2	18.76	91.0	21.46

Table 2 reports means and standard errors of logit difficulty estimates derived from separate Rasch Model calibrations for each group. By this procedure negative difficulty estimates indicate comparative ease, while positive difficulty estimates indicate comparative difficulty of the items in the respective subtests. Note that the BICAL procedure sets mean item difficulty for the total test at zero, so that total test differences are obscured.

Table 2 Mean Rasch difficulty estimates and standard errors for two language groups on five subtests and ESLPE total

Subtest	Chinese (N = 76)		Spanish (N = 33)	
	$\bar{X}$	s	$\bar{X}$	s
Listening	0.42	1.50	0.40	1.71
Reading	−0.56	1.19	−0.49	1.20
Grammar	−1.25	2.00	−0.56	1.16
Vocabulary	0.11	1.62	−0.70	2.11
Writing error detection	0.88	1.24	0.74	1.23
Total	0.00	1.56	0.00	1.39

Note that the results of Table 2 parallel those of Table 1, with vocabulary items appearing to comprise the subtest with greatest advantage for Spanish speakers, while the grammar items seem to reflect the greatest comparative advantage for Chinese speakers. Since mean total scores are arbitrarily set at zero, and since there is reason to suspect that the population invariance hypothesis is not fully met with this data set, it is not readily possible to make absolute comparisons of these means across subtests; nevertheless, it is confirmatory to note the same trends as were visible in Table 1.

Table 3 reports the Pearson correlations between the difficulty calibrations for the two groups on total test and all subtests.

Note from Table 3 that the weakest relationship between difficulty estimates for the two language groups was present in the subtest

Table 3 Pearson product-moment correlations, regression intercepts, slopes, and standard errors of estimate for the relationships between Rasch difficulty calibrations of two language groups

Subtest	*n*	*r*	*a*	*b*	*s*
Listening	30	0.89	−0.02	1.01	0.80
Reading	30	0.73	−0.07	0.74	0.83
Grammar	30	0.74	−0.02	0.43	0.79
Vocabulary	30	0.31	−0.74	0.40	2.05
Writing error detection	30	0.76	0.07	0.75	0.81
Total	150	0.64	−0.07	0.61	1.25

of vocabulary. This finding further indicates irregularities associated with vocabulary item difficulties estimated for the two groups. This suggests that the vocabulary subtest was the greatest single source of Chinese/Spanish bias for the ESLPE.

Figure 1 provides the actual scattergram and regression line associated with the total score correlation reported in Table 3. A 95 per cent confidence interval was constructed around the regression line. Items falling above the confidence interval showed bias in favour of the Chinese speakers, while items below the confidence interval displayed bias in favour of the Spanish speakers.

As Figure 1 indicates, only four items were found to lie outside the confidence interval. Each one of these items was biased in favour of the Spanish speakers. On closer inspection all four items were found to come from the vocabulary recognition section. All four English items were found to possess close cognate forms in the Spanish language. The biased vocabulary items along with their Spanish cognates were the following: approximate (*aproximado*), animated (*animado*), maintain (*mantener*), obstruct (*obstruir*).

A different procedure might have been followed for the determination of the confidence interval. The standard errors for the two difficulty estimates could be pooled in the manner described by Wright and Stone (1979) for each data point. The resultant confidence interval would be curvilinear and narrower near the means. In this particular case it is doubtful that this procedure would have identified different items as biased than those which were identified by least-squares regression.

A further technique was employed as a possible means of detecting item bias. Rasch Model total fit *t*-values were compared for the two language groups and across all five subtests. However, since no more than five out of 150 items showed misfit for either language group, and since there was no clear pattern across subtests, that approach was abandoned as unfruitful.

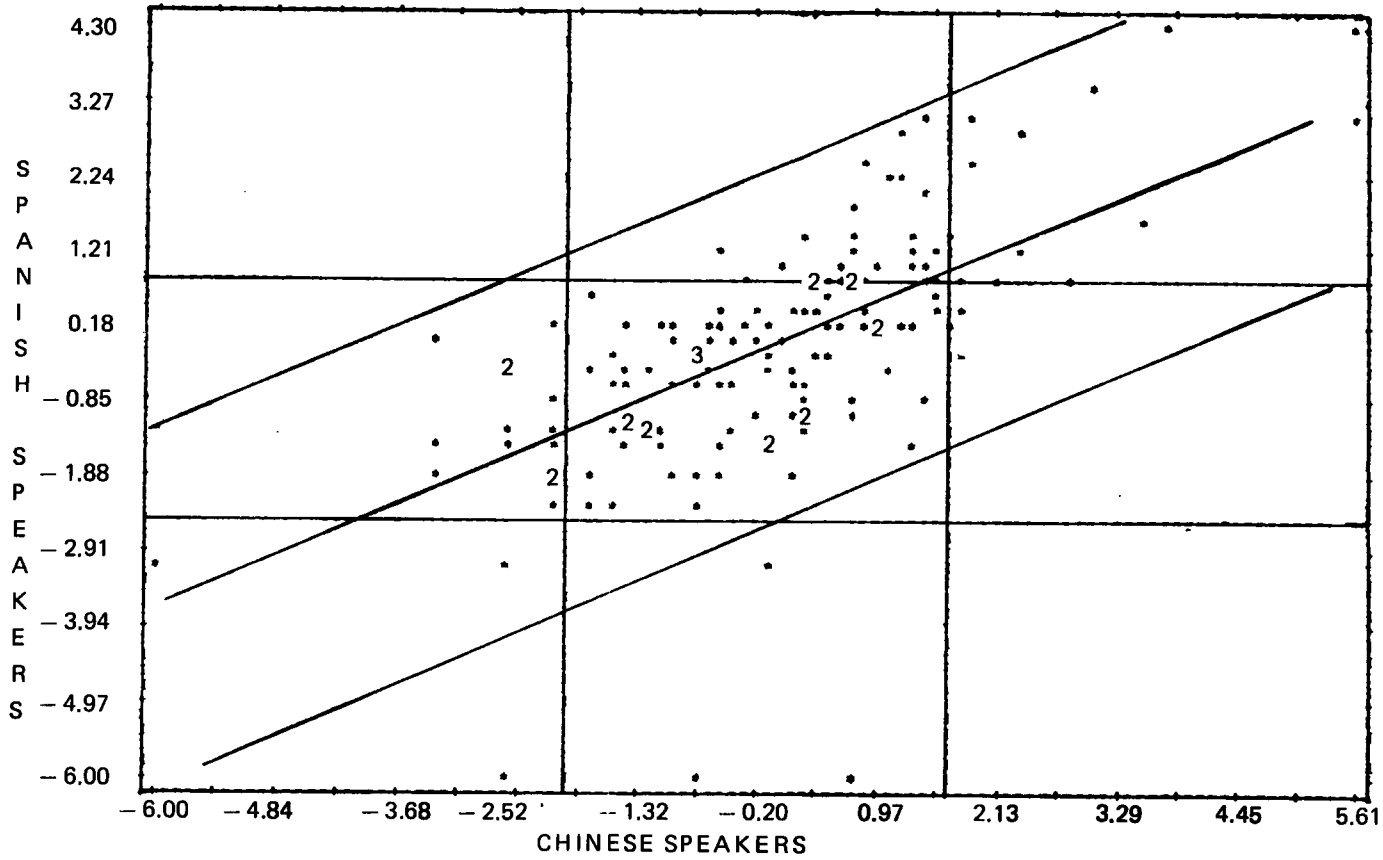

Figure 1 Regression of ESLPE logit difficulty estimates for Chinese speakers on those for Spanish speakers with 150 ESLPE items

III Discussion

Examination of 150 ESLPE items for Chinese/Spanish native speaker bias revealed only four vocabulary recognition items which appeared to exhibit bias according to the criteria which had been established. Each of the vocabulary items and many of the distractors in those items were found to consist of English words for which close cognate forms existed in the Spanish language which were not similarly available in Chinese.

Several interesting questions stem from the findings reported here.

1) Is a 95 per cent confidence interval an appropriate standard for the determination of bias? Had a narrower interval been set, additional items would have been identified as biased. Probably the answer to this question would centre around the certainty with which one would wish to assert that bias were present, or the certainty needed to affirm that bias were not present. Setting a narrower confidence interval for the identification and elimination of biased items would result in a more rigorous standard and the greater likelihood that bias would be eliminated. The present procedure would permit more confident generalization that the four identified items were truly biased.

2) Another important question has to do with the consequences of the detection of statistical bias. Should we now eliminate the four vocabulary items since they are clearly biased in favour of the Spanish speakers? At least two responses are possible to this question. First, it may be argued that correct responses to those items on the part of the Spanish speakers do not reflect knowledge of English on their part, and therefore the items should be eliminated. Another, perhaps more logical, position is that the advantage that accrues to Spanish speakers on these items stems from the greater similarity of their language to English. Presumably the two languages share lexical content that would make it easier for the native speakers of Spanish to recognize English vocabulary items. Because of these similarities, not only would Spanish speakers be expected to do better on English vocabulary tests, but they might also be more easily able to master the English lexical system as a whole. This is another way of saying that any item bias present on the test may be representative of some underlying content bias in the target language. The position that might derive from these considerations is that, as long as the lexical items appearing on the test are randomly and broadly sampled so that they can be said to represent the lexicon of the target language validly, and since the goal of the

examinees is to demonstrate or acquire proficiency in the target language, then there is little concern from a fairness point of view over advantages or disadvantages that occur for subgroup members whose language is consistently more similar or dissimilar than some other language with respect to the target language. Bias, in this view, would be present if a disproportionate number of lexical items were employed for which Spanish cognates existed. As long as the proportion of such items included in the test does not exceed the proportion existing naturally in the languages, test bias would not be present even though Spanish speakers had a natural advantage with a representative number of items. Random sampling of large item domains would best neutralize possible test bias simultaneously across all disparate language groups by ensuring that the proportion of lexical items 'biased' for any language group would be proportionately representative of the content bias between English and the respective background language considered. Presumably the person who knew the most languages similar to English would have an advantage on the vocabulary portion of the English test over other persons, other variables being equal, but this advantage would not be construed as test bias provided the items were randomly and broadly sampled.

IV References

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

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

Cronbach, L.J. 1976: Equity in selection – where psychometrics and political philosophy meet. *Journal of Educational Measurement* 13, 31–41.

Fallis, B., Henning, G. and **Huang, J.** 1985: Rasch model equating and equivalency for language tests. Forthcoming.

Hambleton, R.K. and **Swaminathan, H.** 1985: *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff.

Henning, G., Hudson, T. and **Turner, J.** 1985: Item response theory and the assumption of unidimensionality for language tests. *Language Testing* 2, 141–54.

Linn, R.L. 1976: In search of fair selection procedures. *Journal of Educational Measurement* 13, 53–58.

Linn, R.L., Levine, M.V., Hastings, C.N. and **Wardrop, J.L.** 1981: An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement* 5, 159–73.

Peterson, N.S. and **Novick, M.R.** 1976: An evaluation of some models for culture-fair selection. *Journal of Educational Measurement* 13, 3–29.

Rasch, G. 1960: *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Scheuneman, J. 1979: A method of assessing bias in test items. *Journal of Educational Measurement* 16, 143–52.

Shephard, L.A., Camilli, G. and **Averill, M.** 1981: Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics* 6, 317–75.

Wright, B.D., Mead, R. and **Draba, R.** 1976: *Detecting and correcting item bias with a logistic response model.* Research Memorandum No. 22, University of Chicago, Statistical Laboratory, Department of Education.

Wright, B.D. and **Stone, M.H.** 1979. *Best test design: Rasch measurement.* Chicago: MESA Press.