DOCUMENT RESUME

ED 331 844 TM 016 339

AUTHOR Gafni, Naomi
TITLE Differential Item Functioning: Performance by Sex on Reading Comprehension Tests.
PUB DATE 91
NOTE 10p.; Paper presented at the Annual Meeting of the Academic Committee for Research on Language Testing (9th, Kiryat Anavim, Israel, April 25-27, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Comparative Testing; Foreign Countries; Higher Education; *Item Bias; Reading Comprehension; *Reading Tests; Secondary Education; *Sex Differences; Test Content
IDENTIFIERS *Israel; *Psychometric Entrance Test (Israel)

ABSTRACT

Items in the verbal (Hebrew and English) sections of the Psychometric Entrance Test (PET) administered for university admission in Israel were studied for differential item functioning (DIF) between the sexes. Analyses were conducted for 4,354 males and 4,901 females taking Form 3 of the PET in April 1984, and 3,786 males and 3,815 females taking Form 17 of the PET in April 1987. Three subtests were examined: (1) verbal reasoning; (2) English; and (3) mathematical reasoning (a control non-verbal test). DIF was determined for the 1984 population through: the weighted sum of the differences between the two groups and across all ability groups; and the root of the mean squared differences as defined above. These two indices and a Mantel-Haenszel chi square test examined DIF for the 1987 group. About one-third of the items in the verbal and mathematics reasoning tests were found to have DIF, but few English subtest items did so. The content of some of the items exhibiting DIF was clearly related to stereotypical perceptions of feminine and masculine areas of interest. Implications for test content are discussed. (SLD)

Differential Item Functioning: Performance by Sex on Reading Comprehension Tests

Naomi Gafni

The National Institute for Testing & Evaluation

Jerusalem, Israel

Paper presented at the 9th ACROLT Symposium, 1990

There is an intense concern for the fairness of educational and psychological tests and for the problem of providing appropriate access to higher education for different groups (e.g., minorities). Selective admissions tests must be carefully scrutinized for fairness. One way to determine the fairness of tests, might be an examination of the test components (the items), usually referred to as analysis of differential item functioning (DIF). An item is considered to function differentially for two groups if, when ability is controlled for, a larger proportion of the members of one group responds correctly to the item. Controlling for ability defines comparability of group; other criteria used to define comparability are: schooling or other measures of relevant experience and membership in other groups (Holland & Thayer, 1988). Once an item is found to function differentially, it may be of interest to examine its characteristics so as to enable detection of DIF before an item is used in a test.

All methods used for examination of DIF assume that the test is an overall fair measure for both groups. They detect those items that function differentially relative to the other items comprising the test. The most convenient and most used method in the seventies described in the literature is the delta plot method (Angoff, 1972; Donlon, 1984). According to this method which does not control for ability, deltas (standardized indices of item difficulty) for the two relevant groups are plotted against each other and those items that turn out to be outliers are marked as manifesting DIF. These items tend to be relatively more difficult for members of one group than for the other.

Stern (1978), using this method, examined DIF for black and white

males and females who were examined by the Test of Standard Written English (a part of the SAT). He found that analogies and reading comprehension items were relatively easier for females while vocabulary and sentence completion items were relatively easier for males. Six analogies were detected as "biased" in favor of females, five of which were characterized by philosophical or aesthetical content. In addition, items accompanying three reading comprehension paragraphs, two of which dealt with psychological differences between men and women, revealed DIF; their respective items favored females. Where items were found to favor males, they usually dealt with science or practical matters, subject matters with which males are considered to be more familiar.

Another method for examination of DIF was suggested by Dorans and Kulick (1983). In this method ability is controlled for by dividing examinees into groups based on their test scores so that only comparable members of the two groups are compared. For each such score group the difference between the proportion of correct answers within the males and the females is computed; the weighted sum of these differences across ability groups is used as a measure of DIF. For some cases where items were marked as manifesting DIF, an explanation could be found by relating the content of the item to the direction of the bias. However, for most of the differentially functioning items no explanation could be found.

A method proposed by Holland (1985) and currently used by ETS is the Mantel-Haenszel (MH) procedure. The procedure is noriterative contingency table method for estimating and testing a common two-factor association parameter in a 2 X 2 X K table (Holland & Thayer, 1988).

The suggested statistic has an approximate chi-square distribution with 1 degree of freedom. It provides a test of "no bias" against the alternative hypothesis of "bias".

The objectives of this study were twofold: 1) to examine items included in the verbal sections (Hebrew and English) of the Psychometric Entrance Test (PET) to the universities in Israel for DIF in the two sex groups; and 2) to try and characterize items exhibiting DIF.

Items included in the verbal sections were of a particular interest because content seems to play an essential role in the process of solving them.

METHOD

Population

The analysis was carried out on two populations: 1) 4354 males and 4901 females taking Form 3 of PET in April, 1984; 2) 3786 males and 3815 females taking Form 17 of PET in April, 1987.

Instruments

Three subtests of PET were examined for DIF in this study: verbal reasoning (40 items), English (50 and 48 items for the two forms respectively), and mathematical reasoning (30 and 35 items for the two forms respectively). The mathematical reasoning test was selected as a control non-verbal test to enable a comparison with the verbal tests.

Procedure

Two indices suggested by Dorans and Kulick (1983) were used to examine DIF for the 1984 population: (1) Dstd - defined as the weighted sum of the differences between the frequencies of correct responses in

the two groups, across all ability groups. Members of both groups were considered to be of the same ability if their raw scores on the subtest were the same. (2) RMWSD - the root of the mean weighted squared differences as defined above. The two measures were computed separately for each of the items composing each subtest.

An item was considered to manifest DIF if for two inter-gender random groups the item's index values exceeded the maximal values of two intra-gender control groups. The intra-gender maximal index values for the 1984 population were also used to determine DIF for the 1987 population.

In addition to the above two indices, a MH chi-square test was used to examine DIF for the 1987 population (Form 17). Results yielded by the two methods were checked for consistency.

RESULTS

Form 3

Verbal Reasoning. The item most clearly "biased" - in favor of males - was a syllogism item dealing with cab-drivers. Two other syllogism items dealing with baking cakes and bread, respectively, were "biased" in favor of females. Two reading comprehension items, one of which dealt with law and the other with geophysics, were found to display DIF in favor of males. Four reading comprehension items favored females: one of these dealt with law, another with geophysics, and the other two with psychology. Items of the graph comprehension type exhibited a tendency to favor males, and syllogisms items tended to favor females.

Mathematical Reasoning. Eight items were found to display DIF in

favor of females: three were algebra items, one was a logical problem, and four were numerical series. Five items manifested DIF in favor of males: two were geometry items, one a percentage problem, one an arithmetical problem, and one a scaling problem.

English. Only two out of the 50 items revealed DIF in favor of males, and dealt with physics and astronomy. One item with DIF in favor of females dealt with art.

Form 17

The results based on the two methods (Dorans and Kulick's Dstd and the Mantel-Haenszel test) were found to be consistent with each other in terms of the relative size of their index values. The MH test yielded more significantly "biased" items than Dorans and Kulick's method. This result is probably an outcome of the fact that the statistic provides us with only an approximation of the chi-square distribution. Therefore, the value selected as a cutoff point should be more conservative (higher) than the value suggested by a chi-square table. Consistent with the Dstd values selected as cutoff points, the MH cutoff point turned out to be MH=25 (as opposed to the assumed critical value of 3.84 for alpha = 0.05).

Verbal Reasoning. Thirteen items were marked as "biased", five of which favored females. Two of these items were analogies and the other three were of the reading comprehension type. The items found to favor males were of the table and graph comprehension types.

Mathematical Reasoning. Ten items were marked as "biased", three of which favored females. One of the three was a numerical series, one was a geometry item and the third a numerical problem. Of the seven items favoring males, three were geometry items, three were of the

"work problem" type, and one was a percentage problem.

English. Three items were detected as manifesting DIF, all of which favored males. One item dealt with a basket-ball game, one with dogs, and the third with language.

DISCUSSION

Differential item functioning is an evidence of another dimension involved in the process of measurement beyond the trait intended to be measured by the test. Items marked as "biased" are not automatically excluded from a test. Before such steps are taken an attempt should be made to characterize them and ensure that results are replicated.

The findings of this study indicated a fairly large degree of variability among the subtests with respect to the proportion of items exhibiting DIF. While about one third of the items included in the verbal and mathematical reasoning tests were found to have DIF, few English items exhibited DIF. English is also the subtest in which the overall difference in performance between males and females was the smallest. This variability might be related to the way the different subtests are constructed. While items for the verbal and mathematical reasoning are selected regardless of their content (content is usually considered to be irrelevant to the process of finding the correct answer of these items), there is more emphasis on content when items are selected for the English subtest. This suggestion was supported by findings related to DIF for the figural reasoning subtest (Gafni & Beller, 1989); these items are not characterized by any content and indeed just one item out of the 27 items was marked as "biased".

Since the examination of DIF was internal and was not carried out

against an external criterion, the only inference possible is within the context of the rest of the items; the average DIF will necessarily be zero, indicating no "bias" of the whole test. Given that the findings on bias in prediction indicate that sometimes the PET scores tend to overestimate male performance on the criterion (grade point average), it is important to examine DIF in individual items. The nature of the items exhibiting DIF may suggest a direction for exploring further the source of bias in prediction where it has been found.

The content of some of the items exhibiting DIF was clearly related to stereotypical perceptions of feminine and masculine areas of interest. Since ability was controlled for, the finding indicates differential relative strengths in the different areas for the two genders. This was especially prominent in verbal and mathematical reasoning (which seem to be less unidimensional than English). It seems desirable to represent the different content areas in a test fairly, even when the content does not bear a direct relationship to what is being required by the item. The different content areas should be carefully pre-specified and should not include contents that are clearly identified as characterizing one group.

The DIF results regarding mathematics items in PET support previous findings (Donlon, 1984), in which numerical series and algebra items tended to favor females, and geometry items tended to favor males. It is worthwhile noting that the results obtained for Form 17 replicated those obtained for Form 13.

REFERENCES

Angoff, W. L., (1972). A technique for investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, Hawaii.

Donlon, F. T., (Ed) (1984). The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examinations Board.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms standardization approach. ETS research report 83-9. Princeton, N.J.: Educational Testing Service.

Gafni, N. & Beller, M. (1989). Assessing use fairness of the Psychometric Entrance Test for the two sex groups. Research Report 95. Jerusalem, Israel: National Institute for Testing and Evaluation.

Holland, P. W. (1985). On the study of Differential Item Performance without IRT. Proceedings of the Military Testing Association, October.

Holland, P. W., & Thayer, D., T. (1988). Differential item performance and the Mantel-Haenszel procedure. In: Test Validity. edited by Wainer, H. and Braun, H., I., pp. 129-143. Lawrence Erlbaum, Hillsdale, New Jersey.

Stern, J. (1978). College Board item bias of the Scholastic Aptitude Test of Standard Written English Forms XSA2/E4. Statistical Report 56-78. Princeton, N.J.: Educational Testing Service.