DOCUMENT RESUME

ED 098 262 TM 004 051

AUTHOR Lockheed-Katz, Marlaine
TITLE Sex Bias in Educational Testing: A Sociologist's Perspective. Research Memorandum No. 74-13.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RM-74-13
PUB DATE Aug 74
NOTE 15p.; Paper presented at the International Symposium on Educational Testing (The Hague, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Evaluation Criteria; *Females; Feminism; Item Analysis; Norms; Predictive Validity; *Sex Discrimination; Sex Stereotypes; Sociology; *Test Bias; *Test Construction

ABSTRACT

Several criteria for assessing bias in educational tests are presented and discussed. These criteria were developed in accordance with basic notions of fairness, equality, and expanded life options for women. In terms of prescriptions for test developers, the criteria are: (1) tests should be constructed of items which contain either no sex references or equal sex references; (2) status of males and females within the test should be equal; (3) item content should not reinforce traditional sex stereotypes. Tests currently in use may be considered biased if: (4) item content in terms of male or female statuses or stereotypes affects the performance of males or females differentially; (5) the test predicts differentially for males and females; (6) the test is normed separately for males and females unless separate norms are used to insure balance in selection; (7) the test is constructed so that female futures may be separated from male futures. (Author/RC)

RM-74-13

RESEARCH MEMORANDUM

SEX BIAS IN EDUCATIONAL TESTING: A SOCIOLOGIST'S PERSPECTIVE

Marlaine Lockheed-Katz

Paper presented at the International Symposium on Educational Testing, The Hague, July 16-19, 1973.

Educational Testing Service
Princeton, New Jersey
August 1974

Sex Bias in Educational Testing: A Sociologist's Perspective

Marlaine Lockheed-Katz
Educational Testing Service

The purpose of education, and hence of educational testing, should be to expand the life options of individuals. In too many cases, however, the life options available to women are limited and rigidly defined. In an effort to improve the status of women internationally, the United Nations General Assembly adopted a "Declaration on the Elimination of Discrimination Against Women" in 1967. This document states that discrimination against women is fundamentally unjust and inconsistent with human dignity; it calls for the abolition of existing laws, customs, regulations and practices which are discriminatory against women, and firmly establishes the principle of equal access to education for women.

Present educational practices, however, still reflect commonly held beliefs about fundamental differences between the sexes. Thus schools may be segregated by sex, offer different curricula for males and females, provide fewer resources for female activities than for male activities, and prepare females for different careers from males.

One major adjunct to education is educational testing, yet on several dimensions educational tests appear to reinforce beliefs about the differences between the sexes. A review of major educational tests suggests that men and women are not presented equally in these tests. Unequal distribution of items between male and females, stereotyped images of both men and women, and separate interpretations of test results for men and women are characteristic of such tests. Such inequities may restrict the life options of women.

The purpose of this paper is to present criteria for the construction and evaluation of tests which expand the options of men and women. These criteria may be applied to evaluate tests administered to any heterogeneous population.

There are seven criteria against which a test may be judged for bias.

1. the actual distribution of test items dealing with male and female actors
2. the status of the males and females within the items
3. the content of items relative to traditional or stereotyped male or female interests or skills
4. the effect of (1), (2), or (3) above on male or female success on any item or items or on the test as a whole
5. the overall predictive validity of the test for males and females with respect to some criteria such as future grades and the use of tests for selection based upon such prediction
6. the use of separate norms for evaluating the test performance of males and females
7. the uses made by counselors and others to predict future occupations, interests or skills of males and females as a result of their test performance, when such predictions separate male futures from female futures.

Each of these criteria will be considered in this paper.

1. The distribution of items dealing with male and female actors.

In a 1973 study funded by the Ford Foundation, Tittle investigated the occurrence of male and female references in 9 different series of tests of academic achievement used for American students from kindergarten through

12th grade. Overall, 29 separate tests were examined. Within each test, all references to males and females were counted. The generic use of the word "man" and "he" was counted separately.

Of the tests scored, all but one contained a higher number of male references than female references. The ratio ranged from 14-1 to slightly less than 1-1. Only 8 of the 29 test batteries examined had less than a 2-1 ratio of male references to female references. The older the age group for whom the test was written, the higher the male/female ratio.

Lockheed-Katz (1973) conducted a similar investigation of male and female references, by item, in eight major college and graduate school entrance examinations. Similar imbalance was found within these tests. Items were coded according to whether they contained no sex reference, male only sex reference, female only sex reference or both male and female sex reference. The ratio of male only items to female only items ranged from 16-1 to 2-1. Only 22 of the 1220 items coded contained references to both men and women. Seventy-five percent of all the items contained no sex reference, but of the remaining 25% more than 4/5 were "male only items."

If tests are to be constructed in such a way as to reflect an implicit egalitarian ideology, then either sex referenced items should be entirely eliminated or a balance between items dealing with male actors and female actors should be achieved.

2. A second criterion for judging a test for bias is to determine how the male and female actors are portrayed ~~within items.~~

In her study of sex-role stereotyping in achievement tests, Tittle reported that "women are portrayed almost exclusively as homemakers or in the pursuit of hobbies." Furthermore she reported that "some items imply that the majority of professions are closed to women."

In the investigation of the college and graduate school admissions tests conducted by Lockheed-Katz, the relative status of males and females in a single item was coded. Of the 22 items (out of 1220) which contained both a male and a female actor, 10 items portrayed the men and women as equal in status; the remaining 12 items showed the men as being higher status than the women. No item on any of the eight tests portrayed a woman in a higher status position than a man, for example, as a female principal with a male teacher or a female lawyer with a male client.

Typically females were referred to as mothers, teachers, secretaries or wives. Males were referred to as lawyers, managers, principals, superintendents, doctors or other professionals. Since it is both the case that there are now women in high status positions relative to men and that true equality of opportunity implies that men and women should have equal access to both high and low status occupations, tests should also reflect this equality.

3. A third criterion against which to judge a test for bias is the distribution of items relative to traditional or stereotyped male or female interests or skills. In American testing this general area of concern, which may be considered the cultural relevance of a test, has been centered about the issue of minority representation on tests. Thus studies by Quirk and Medley (1972), Linn (1973) and others suggest that American tests have been culturally specific and may be inappropriate for use with non-Anglo populations.

The same may be said with respect to the male-ness of tests and items. Unfortunately, the test constructor may face a dilemma: to include items

which reflect "female" interests may imply the reinforcement of sex-role stereotypes. That is, to include kitchen measurement items in a math test may make the item more manageable for females at the cost of reinforcing the stereotype which says that woman's place is in the home.

It is possible to construct items in which the actor and the actions are not stereotypically associated. Such items might include a boy measuring flour in the kitchen or a girl fixing a bicycle. The culturally specific item is retained, but the stereotype is broken by substituting a different actor than would be expected. No studies have been found which report any but extreme sex role stereotypes in test items.

The preceding three criteria against which to judge a test

--test items to be balanced with respect to male and female references

--the relative status of males and females within items to be balanced of equal

--cultural interests of males and females to be balanced with reinforcing cultural stereotypes

are based upon the layman's notion of fairness, which implies an equal representation of conflicting or divergent interests or likes. If a test is biased according to these criteria, it is not intrinsically fair as it does not represent males and females equally, whether or not the test discriminates between male and female test takers.

4. Another criterion for assessing bias is the ability of either items or a total test to discriminate between males and females. In other words, what effect do the three previously mentioned biases have on the performance of males and females on tests?

At present, there is little reported that answers this question. Some studies report analyses of these issues in connection with ethnic diversity.

Echternacht, Carlson and Flaugher (1973) described three studies examining differences in item difficulty for black and white test takers. They describe alternative strategies for assessing test bias. The first strategy was to regress the Δ, or inde· of item difficulty, for blacks or whites for each item type on a give· est. This allowed the investigators to determine if certain types of items were more difficult for blacks than for whites.

A second approach employed by Echternacht, Carlson and Flaugher was to determine if, for any pair of items, the easier of the items for the black test takers was the harder for the white and vice versa. By using this paired comparison, a test could be constructed that would be equally difficult for blacks and whites, although the difficulty would differ at the item level.

A third strategy suggested was to correlate black and white item responses to items, producing average within race and cross race correlations and to compute ANOVA on these correlations to determine if item-race interactions existed.

Echternacht (1972) reported a study of item-sex interaction, using the first method described above, that found 3 items out of 30 in the Aptitude Test for Graduate School of Business which showed differences between males and females in response; no information about the nature of the items was included in the report, however.

Another study of total test performance for men vs. women was conducted by Swineford (1972) on the Law School Admissions Test. Two kinds of items

showed differences in performance. Women performed better on verbal items while men performed better on Data Interpretation sections. No attempt was made to examine the items individually.

Coffman (1961) examined items in an aptitude test and made predictions with respect to the content of the items as to which items would favor men and which would favor women. These items were judged for traditional interest or skills of men and women. Of the 16 judgments, 14 were in the predicted direction. Of the nine items which involved mechanical knowledge, science or business and were judged to be easier for men, eight actually were easier for male respondents. Of the ten items which involved personal feelings or personality characteristics and were judged easier for women, nine were actually easier for women.

Donlon (1971) reported a study patterned after Coffman's in which the scores of the 103,275 persons who took the Scholastic Aptitude Test in May 1964 were examined for sex differences by item. Items for which the women's performance was statistically superior to the men's, and items for which the men's performance was superior to the women's were located and examined for sex related bias. Of the 90 verbal items on the test, 8 items favored men and 11 favored women. Seven of the eight items favoring males were coded as having scientific or "practical affairs" content; 8 of the 11 items favoring females were coded as having human relations, humanities or aesthetic-philosophical content. The content of these items, relating to stereotyped or culturally associated interests of men and women, apparently accounted for the differences in male and female performances on the item.

Milton (1958) reported five studies in which differences between males and females in problem solving were explained by the sex-role stereotype of

the problem. Milton documented and replicated the finding that when problems were framed so as to make them less appropriate to the masculine role, sex differences in problem solving were reduced.

5. A fifth criterion for judging whether a test is biased relates to its overall predictive validity with regard to an external criteria, such as future grades. Typically this sort of test analysis is used for selection of test takers into college or occupation. In fact, it appears difficult to separate the issues of predictive validity of a test from the use of the test for selection purposes. Tittle's review of test predictive validity begins by summarizing research conducted on black-white and male-female predictions of college grades from Scholastic Aptitude Test scores, and ends by reviewing models of selection bias.

Seashore (1962) summarized several studies which conclude that women's grades are more accurately predicted by tests than are men's. These conclusions are reiterated by Cole (1973) who analyzed data from students enrolled in 19 American coeducational colleges; he found that standard test and high school grades better predicted women's first term grades than men's first term grades.

The main rationale for examining the predictive validity of a test, however, is for selection purposes. Selection is itself subject to bias.

Cole (1972) distinguishes six models of selection bias and applies these models to male-female selection. A single set of data is analyzed according to the six models, and judgments about the fairness of the selection practice are made based upon an analysis of the data. The six models of selection bias are:

1. The regression model, where test bias refers to either over-prediction or underprediction of criterion measures for different populations using a single regression equation.

2. The quota model, where selection is made according to the percent distribution of each subgroup within the total population.
3. The subjective regression model, in which a constant is added to minority group scores to increase the probability of their selection according to the regression model.
4. The equal risk model, where all persons who have the same probability of being successful on a criterion measure are selected.
5. The constant ratio model, where selection by group is made in proportion to that group's success on the test.
6. The conditional probability model, in which the probability of being selected is contingent upon achieving a satisfactory criterion score and is not related to group membership.

Applying these models to the selection of men and women into college, ACT (1973) reported that using separate regression equations in the regression models is fair, but that combined equations are biased against women. Linn (1973) also reported a similar finding with regard to black and white male and female students accepted at 22 different colleges. That is, women's achievement is underpredicted using a regression equation based upon all male or combined data.

The other models of selection bias applied by Cole to the 19 school data revealed different patterns of fairness to men and women.

The quota model is frequently preferred for the selection of men and women as it permits the selection of fewer women than would qualify under the regression model.

Both the constant ratio and the conditional probability models were found to be unfair to men, while the equal risk model was judged to be fair but impractical.

6. A sixth criterion for judging bias in a test centers about the implicit assumptions associated with reporting separate norms. When norms on a test are presented separately for males and females, as they are for many tests, it reinforces the implicit belief that male test performance and female test performance should be different. The reinforcement of beliefs regarding male and female abilities by separate norming is an issue which is distinct from the issue regarding the use of norms for selection purposes.

Although the normal distribution of test scores may at present be different for males and females, there is some evidence that cultural rather than genetic influences account for these differences.

For example a study by Fremer, Coffman, and Taylor (1968), "The College Board Scholastic Aptitude Test as a predictor of academic achievement in secondary schools in England," showed that while American freshman girls score higher on verbal than on mathematical aptitude, British girls of the same age score higher on the mathematical than the verbal tests. Peck (1971) also reported no notable systematic sex differences in performance on aptitude and achievement tests across eight different countries. He attributed such differences as occur to cultural differences.

When norms are used for selection purposes, however, it may be that separate norms will promote more equalitarian representation of males and females in the roles for which they are being selected. Thus, selecting on the basis of the top 10 percent of the males and the top 10 percent of the females will yield a selected group balanced by sex, while selecting on the

top 10 percent of a combined group of males and females will yield a selected group which is imbalanced toward one or the other sex if the norms of the two groups are actually different.

The issue of separate norming is consequently complicated. While reporting separate norms may reinforce stereotyped beliefs regarding male and female abilities, failing to do so may reduce the likelihood of equal male or female selection.

7. The final criterion for evaluating a test for bias is the extent to which it may be used to separate male futures from female futures. This bias is most clearly observed in vocational interest tests which typically separate male and female interests.

For example, Tittle reported that the Strong-Campbell Interest Inventory, a unisex modification of the Strong Vocational Interest Blank, contains occupations identified by the letter "m" or "f" for male or female. Although physician is listed as an occupation for both males and females, biologist, cartographer, social scientist, architect, minister, school superintendent and sales manager are identified only as male occupations. The Kuder Occupational Interest Survey contains even more blatant distinctions between male and female occupational and educational interests. Tittle reported of the 77 male occupational scales and the 57 female occupational scales, there are only 16 scales for identically stated occupations for males and females. College major scales also reflect separate and unequal opportunities for men and women. The historical fact that women did not have, or were not allowed to have, certain interests is surely no excuse to discourage future generations of women from considering these interests.

This paper has attempted to present briefly several criteria for assessing bias in educational tests. These criteria were developed in accordance with basic notions of fairness, equality and expanded life options for women. To summarize these criteria in terms of prescriptions for test developers, they are:

1. tests should be constructed of items which contain either no sex references or which are balanced for male and female references
2. the status of the males and females within the test should be equal
3. the content of items should not reinforce traditional or stereotyped images of men and women.

Tests which are currently in use may be considered biased if:

4. the content of the items in terms of male or female statuses or stereotypes effects the performance of males or females differentially
5. the test predicts differentially for males and females
6. the test is normed separately for males and females unless separate norms are used to insure balance in selection
7. the test is constructed so that female futures may be separated from male futures.

The principle of fairness which calls for eliminating discrimination against women and providing women with equal access to education requires that all aspects of education be free from discriminatory material. This requirement applies to educational tests in particular, as tests of achievement and aptitude typically determine both men and women's access to future education.

References

American Council on Testing.
1973 "Assessing students on the way to college." American Council on Education Technical Report, ACT Publications, Vol. 1, Ch. VII.

Coffman, W. E.
1961 "Sex differences in responses to items in an aptitude test." Eighteenth Yearbook, National Council of Measurement in Education: 11, 7-124.

Cole, N. S.
1972 "Bias in selection." ACT Research Report No. 51. Iowa City, Iowa.

Cole, N.S.
1973 "A model for fairness in selection." Paper presented at the American Educational Research Association Annual Meeting, New Orleans.

Donlon, T. F.
1971 "Content factors in sex differences on test questions." Paper presented at the meeting of the New England Educational Research Organization, Boston. (Also Research Memorandum 73-28, Educational Testing Service, Princeton, N.J.)

Echternacht, Gary J.
1972 "An examination of differential item response characteristics for six ATGSB candidate groups." Project Report 72-4, Educational Testing Service, Princeton, N.J.

Echternacht, G. J., Carlson, A. B., and Flaugher, R. L.
1973 "Differences in test and item performance for black and white undergraduates." Educational Testing Service, Princeton, N.J., (also published as GREB No. 70-8).

Fremer, J., Coffman, W., and Taylor, P. H.
1968 "The College Board Scholastic Aptitude Test as a predictor of academic achievement in secondary schools in England." Journal of Educational Measurement, Vol. 5, No. 3, 235-241.

Linn, R. L.
1973 "Fair test use in selection." Review of Educational Research, Vol. 43, No. 2, 139-161.

Lockheed-Katz, Marlaine E.
1973 "Test bias in ETS tests." Unpublished memorandum, Educational Testing Service, Princeton, N.J.

Milton, G. A.
1958 "Five studies of the relation between sex-role identification in problem solving." Technical Report 3, Yale University, New Haven, Conn.

Peck, R. F.
1971 "A Cross-national comparison of sex and socio-economic differences in aptitude and achievement." U.S. Office of Education, Washington, D.C. (ERIC Abstract, 049 315).

Quirk, T. J. and Medley, D. M.
1972 "Race and subject-matter influences on performance on general education items of the National Teacher Examinations." Research Bulletin 72-43, Educational Testing Service, Princeton, N.J.

Seashore, Harold G.
1961 "Women are more predictable than men." Journal of Counseling Psychology, Vol. 9, No. 3, 261-277.

Swineford, Frances
1972 "Law School Admissions Test. Comparisons of white male candidates with white female candidates." LSAT Research Summary, Educational Testing Service, Princeton, N.J.

Tittle, Carol K.
1973 "Women and educational testing: A selective review of the research literature and testing practices." The Ford Foundation Division of Education and Research, New York.