DOCUMENT RESUME

ED 129 886 TM 005 689

AUTHOR Dwyer, Carol Anne
TITLE Test Content in Mathematics and Science: The Consideration of Sex.
PUB DATE [Apr 76]
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Item Analysis; *Mathematics Education; *Science Education; *Sex Differences; Sex Discrimination; Sex Stereotypes; *Test Bias; *Test Construction

ABSTRACT

A traditional generalization is that girls are superior in verbal skills and boys in mathematics and the sciences. But most specialists in this area now concede that there is almost more exception than rule in this generalization, and that individual test items may actually modify observed patterns of sex differences. Sex role stereotyping and the issue of male/female representation in test content have often been glossed over with respect to mathematics tests, and, to a lesser extent, with respect to science tests. The effects of item type and item context on sex-differentiated performance are better documented. The balancing of these two aspects of test content is important to remember in the construction of tests. It is also important to have a close match between the test item content and the curriculum or aptitude areas they are intended to measure. There are several sets of useful guidelines available for eliminating sexist content in these materials, but developers should be aware that such efforts cannot be expected to influence test performance for either sex. The issue of performance-related test content must remain a completely separate one, to be resolved in psychometric rather than value-oriented terms. (Author/BW)

TEST CONTENT IN MATHEMATICS AND SCIENCE:

THE CONSIDERATION OF SEX

by

Carol Anne Dwyer

Educational Testing Service

Princeton, New Jersey

A paper presented at the

American Educational Research Association

San Francisco, California

April, 1976

Introduction

It has often been the case, expecially in recent years, that the correlates of sex differences in mathematics and science achievement have been examined, rather than the basic premises themselves.

The traditional generalization is that girls are superior in verbal skills and boys in mathematics and the sciences. But most specialists in this area now concede that there is almost more exception than rule in these generalizations. There are variations in patterns of sex differences associated with age, ethnicity, socioeconomic status, and ability level; yet a sort of halo effect still surrounds mathematics and science achievement and seems to affect researchers as well as teachers and the general public. Fennema (1974) has pointed to a tendency among researchers to report results that conform to a prior expectation (i.e., that males will do better than females in mathematics), rather than results that conform to the actual data they have collected. Data showing superior performance by females may be ignored when authors reach their conclusions, while weaker data showing superior performance by males will be stressed. A recent issue of the newsletter of the National Assessment of Educational Progress (1975) provides an interesting illustration of this tendency. The headline of the newsletter says, in bold letters, "ADULT MALES ON PLUS-SIDE IN MATH BASICS." The article itself, however, indicates that among 13-year-olds, females are superior and there is no sex difference among 9- and 17-year-olds. Of the four age groups studied it is only among adults that male performance is higher. The headline thus does not seem to represent a logical summary conclusion abstracted from the data presented, but rather a response influenced by traditional notions of sex-difference patterns.

A mind-set such as this is most difficult to alter, but I hope my presentation today may contribute to replacing it with more logical deductions by pointing out how such concrete and easily manipulated factors as individual test items may actually modify observed patterns of sex differences.

There are several ways of considering test content that must be taken into account if we are to consider the full range of implications of sex-related achievement data. There are, however, two major kinds of test content which must be kept clearly separate for data interpretations and value interpretations to be meaningful. The first is a set of questions concerning sex-role stereotyping and sexism in test content. This includes such factors as the proportionate representation of males and females in test content, and the roles they occupy. The second is the content of the test in terms of item types, item context, and technical test specifications. These latter issues are all straightforward psychometric considerations.

I would like to stress that the first group of factors, those concerned with sex-role stereotyping and sexism, should be dealt with primarily in terms of social policy and values. Only the second group of content considerations is currently suited to interpretation by psychometric analysis.

Sex-role Stereotyping and Sexism

Work by Tittle et al. (1974) and Lockheed (1973) has extensively documented the occurrence of sex role stereotyping in tests ranging from the primary level through graduate school. This stereotyping has been documented with regard to the questions of the proportionate representation of males and females in test items and to the number and quality of roles in which males and females are portrayed in the test items.

This issue of male/female representation in test content has often been glossed over with respect to mathematics tests, and, to a lesser extent, with respect to science tests. Clearly there are certain large classes of items in these tests where human beings, their interests, and their activities, are not referred to at all. However, there are other large classes of mathematics and science test items that are concerned with human beings, and these items show the same proportionate representation of males and females as do tests of reading or literature -- male references far outnumber female references, and females are typically portrayed in only a limited and very traditional array of roles. There has been very little research effort aimed at establishing a connection between this kind of sex-role stereotyping and test performance.

This dearth of conclusive research evidence should not be interpreted as suggesting that these sexist practices are without effect. As Tittle (1974) has pointed out, educational measurement instruments may reinforce sex-role stereotypes and restrict the range of career choices available to members of both sexes. Lockheed (1973) has also commented that a test is not intrinsically fair if it does not represent males and females equally, whether or not the content affects performance by females and males differentially.

Test Specifications

Other types of test content, which comprise typical elements of test specifications, have been more clearly related to sex differentiated test performance. These include test item types measuring different parts of mathematics or science curricula (for example, geometry or algebra or biology); the cognitive skills required of the test taker (for example, computation, application, analysis); and the sex-relevant context in which test items are set (for example, is a math or science problem set in a context of stock brokerage or home-making or football?). There is also the related issue of the psychometric properties of a test which may be related to the sex of the test taker, such as speededness.

Item Type and Skills

Most tests of mathematics and science begin with a set of specifications which prescribe, in varying degrees of detail, the curriculum content and the learner skills to be covered by the test material. These specifications are very global for most mathematics and science achievement tests. A test may simply specify 20% "Application" items. These items may require application of mathematics (or science) knowledge to a wide range of problems, covering content in very diverse areas of the curriculum (which may or may not be sex-differentiated in performance), and set in unspecified types of sex-related context.

When these tests are then used as the basis for making statements concerning sex differences in test performance, the variations in test content must first be determined, and then be taken into account when the test results are interpreted.

As a specific example, two survey tests of mathematics achievement may differ widely in the proportion of algebra items included in each. There is some evidence (Donlon, 1973) that females do relatively better on algebra items than on other types of test items. The proportion of algebra items included in a test, then, can be expected to influence the test performance of females relative to males: when more algebra items are used, females' scores can be expected to rise; when fewer are used, their scores will drop.

Donlon (1973), in the same study of content factors related to sex differences in the Scholastic Aptitude Test, also concluded that "the approximate 40 point difference between the sexes on this test in scaled scores is, at least in part, a function of the test specifications...if the items were limited to 'algebra,' the difference could diminish to about 20 points. (p. 16)"

Item Context

The context of test items has also been found to affect test performance by males and females in many subject matter areas. In tests of verbal ability, for example, all other things being equal, males achieve higher scores than females when the material to be read or evaluated is set in a context of business, science, practical affairs, mechanical principles, or mathematics.

It is interesting to note that the area of science has been treated globally in this research: no distinction has been made among the various branches of science, even though there is some evidence that females may have higher achievement in biology than in other areas of science. Females achieve higher scores on material drawn from the arts and humanities, or based on understanding human relations. Similar work done in England (King, 1959) has shown the same relative advantage for males with practical and scientific reading passages. These categories to correspond to traditional conceptions of sex roles. There is, however, no conclusive evidence as to whether these context-related sex differences are a result of familiarity with the context, or of motivational considerations associated with the context, or some combination of the two.

Related research by Milton (1957), in the area of mathematical word problem-solving has indicated that test performance may be a function of the sex-appropriateness of the item content. Later work by Hoffman and Maier (1966) failed to replicate this finding, but it is interesting to note what the context was for the so-called "masculine" and "feminine" versions of the problems to be solved in their study. In one problem set, the "feminine" version involved dieting and trying on a new dress; the "masculine" version involved a snail crawling up a wall and slipping back again. It is difficult to understand how this latter example could be construed as "masculine", even in the broadest and most stereotyped sense of the word, which renders interpretation of the study's results somewhat more difficult.

A limited amount of research on the effects of mathematics item context has also been shown a relationship between sex-related contexts and item difficulty. One study by Coffman (1961) examined items in the College Board Scholastic Aptitude Test's quantitative section and made predictions, based of judges' perceptions of the traditional interests or activities of men and women, as to which items would be easier for males and which would be easier for females. For 14 of the 17 judgements made, item data were found to be in the predicted direction ($p > .006$). Of the seven items judged to be easier for males, six actually were easier for them (an interesting point is that five of these "masculine" items involved science content). Of the ten items which were judged easier for females, eight actually were easier for them.

Work done by Strassberg-Rosenberg and Donlon (1975), using the method of delta-plots, also confirms the finding that items which are biased in favor of males tend to be those items having content representing traditional male activities and interests, which include mathematics and science. Conversely, items which are biased in favor of females are those items having content representative of traditional female activities and interests.

Test Balancing

Test specifications for tests of verbal ability often prescribe the numbers and proportions of various item contexts to be included in a test, but this is not typically done for mathematics tests. In the case of verbal tests, where females have typically performed better than males, test specifications require inclusion of material on which males might be expected to perform better, such as material with a science context. The same has not typically been done for mathematics tests, where males' performance is often higher than that of females. I believe that this is an example of a nonconscious sexism on the part of test-makers and researchers: the need for balancing is more readily perceived where males are at a disadvantage, and steps are taken to remedy the condition. No parallel steps are taken when it is the females who are at a disadvantage, perhaps because this situation seems so familiar to all of us that questions about it do not readily come to mind.

Other instances come to mind of test construction procedures in mathematics which are not responsive to the issue of the relationship between test content and differential performance by females and males. The Scholastic Aptitude Test recently dropped the data sufficiency item type from its Quantitative section. This was an item type on which females' performance had been relatively high compared to their performance on other item types. Dropping this type was done, after long and careful deliberation, on the basis of a number of considerations. Data sufficiency items required more reading in their directions and in their texts than other item types and therefore included a larger proportion of verbal skill testing than other item types. Data sufficiency items were also disproportionately difficult fo minority students and contributed to lowered test reliability for minority groups. The first consideration, the involvement of verbal skills, seems to me insufficient reason for eliminating an item type. Real-life performance of tasks in mathematics and science, as well as school learning in these curricula, require skills in responding to both verbal and spatial stimuli. These skills which are probably sex-differentiated, might better be considered reflections of the practical demands of the actual tasks than seen as contaminations of a hypothetical pure mathematical factor.

The second consideration with regard to data sufficiency, the relative difficulty of the item type for minority group members, seems an adequate and important reason for altering test specifications, and one that would be of great interest to researchers making ethnic group comparisons on measures of mathematical aptitude. But the possible effect of this content change on patterns of sex differences in mathematics was not specifically considered. The overall effect of this change (which involves many items in the Quantitative sections) may well be to increase the male-female difference in the mathematics score. Certainly an argument could be made against making a content change that could serve to reinforce a sex-role stereotype. But researchers who have been using the SAT to track changes in sex differences in mathematics should also consider the potential effect of a test content change such as I have described when interpreting the data based on an instrument whose content has been changed, and when drawing their conclusions from this data. In practice, knowledge of such changes may be difficult for many researchers to obtain. For reasons of test security or other internal considerations, content changes may not be announced publically and test booklets covering a period of years may not be readily available for inspection.

Another instance of test content changes that may bear on the interpretation of sex difference data is the content shift that occurs with increasing grade or age level, from a preponderence mathematics computation items to more emphasis on mathematical concepts and problem-solving. This content shift influences interpretation of sex difference results, since tests of mathematics computation typically show no sex difference at any age level. Tests of mathematics concepts tend to show sex differences beginning at adolescence, but concepts are less often measured before this age level (particularly on tests whose specifications were set after the mid-1960's). In using a multi-level mathematics test series in a study with a cross-sectional design, a researcher should assure herself that any test content which might be sex-differentiated has remained comparable throughout the age span if she wishes to make cross-age comparisons of sex differences or to draw developmental inferences. This practice would also have the effect on insuring a better match between test content and curriculum at the elementary school age levels. Elementary school mathematics educators have severely criticized many norm-referenced tests for their heavy dependence on computational behaviors, on the grounds that this content does not adequately sample the realities of elementary and secondary classroom mathematics behaviors (Carry, 1974).

The question of computational versus conceptual test content may also have relevance for the interpretation of longitudinal sex difference data, since test content in the past fifteen years or so has increasingly shifted away from a computational emphasis toward a more conceptual approach. This has been particularly true for the upper grades. But it seems that we now are entering a new phase of the historical cycle which may be characterized as a "back to the basics" movement. There is currently a great deal of test development work being done in the area of "minimal competencies" or "functional literacy" in both reading and mathematics. The demand which impelled this effort may eventually find expression in a return to greater computational emphasis on a broad spectrum of mathematics tests, thus completing another cycle of the computation/concepts shift.

There has also been a recent trend, in a wide range of subject matter areas, away from speeded tests. This is probably in part a reflection of interest in setting educational objectives and assessing mastery in terms of criterion-referenced measures. There has been concern voiced by minority representatives and by women that speeded tests may be penalizing them. A small study by Graf and Riddell (1972) supports this assertion as far as women are concerned. Graf and Riddell point out that most tests of quantitative ability are tests of speed as well as power and will therefore discriminate against females. They further suggest, on the basis of their research, that "one could significantly decrease between sex differences in [mathematical] problem-solving by giving power tests rather than tests which rely heavily on speed [p. 452].

The test content factors I have discussed should all be examined by the interested reséarcher or teacher at the item level. Test specifications tend to be written in very general terms, and, since sex differences are often small, one or two biased items may make a significant contribution to a reported sex difference. If sex differences are to be assessed, even incidentally, it is important to have a close match between the test items' content and the curriculum or aptitude areas they are intended to measure. This will enable the researcher to avoid overgeneralizations concerning sex related performance in an undifferentiated area labeled "mathematics."

In the recent past, educators and others have become increasingly concerned with the inequitable presentation of the sexes in tests and other curriculum materials. There are several sets of useful guidelines available for eliminating sexist content in these materials, but developers should be aware that although this is an important step in test construction, we cannot expect such efforts to influence test performance for either sex. The issue of performance-related test content and context must remain a completely separate one, to be resolved on its own merits, in psychometric rather than value-oriented terms.

REFERENCES

Anonymous. Adult males on plus-side in math basics. NAEP Newsletter, 1975, 8, 1.

Carry, L. R. A critical assessment of published tests for elementary school mathematics. The Arithmetic Teacher, 1974, 21, 14-18.

Coffman, W. E. Sex differences in responses to items in an aptitude test. Eighteenth Yearbook, National Council on Measurement in Education, 1961, 117-124.

Donlon, T. F. Content factors in sex differences on test questions. Research Memorandum 73-28. Princeton, N. J.: Educational Testing Service, 1973.

Fennema, E. Mathematics learning and the sexes: a review. Journal for Research in Mathematics Education, 1974, 5, 126-139.

Graf, R. Q. & Riddell, J. C. Sex differences in problem-solving as a function of problem context. Journal of Educational Research, 1972, 65, 451-452.

Hoffman, L. R. & Maier, N. R. F. Social factors influencing problem solving in women. Journal of Personality and Social Psychology, 1966, 4, 382-390.

Lockheed, M. Sex bias in educational testing: a sociologist's perspective. Paper presented at the International Symposium on Educational Testing, The Hague, 1973.

Milton, G. A. The effects of sex role indentification upon problem solving skill. Journal of Abnormal and Social Psychology, 1957, 55, 208-213.

Milton, G. A. Sex differences in problem solving as a function of role appropriateness of problem content. Psychological Reports, 1959, 5, 705-708.

Strassberg-Rosenberg, B. & Donlon, T. F. Content influences on sex differences in performance on aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D. C., 1975.

Tittle, C. K., McCarthy, K., & Steckler, J. F. Women and Educational Testing. Princeton, N. J.: Educational Testing Service, 1974.