DOCUMENT RESUME

ED 247 237 TM 840 381

AUTHOR Doolittle, Allen E.
TITLE Interpretation of Differential Item Performance Accompanied by Gender Differences in Academic Background.
PUB DATE Apr 84
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Higher Education; *Item Analysis; Mathematics Achievement; Research Design; *Sex Differences; *Test Bias; Test Items
IDENTIFIERS ACT Assessment; *Differential Item Performance

ABSTRACT

The definition of differential item performance (DIP), often referred to as item bias, is discussed. DIP is suggested as a comprehensive term to encompass item bias (item invalidity which is unfair to certain population subgroups) and instructional bias (a valid reflection of group differences in instruction or background). This study investigated the plausibility of an instructional bias interpretation of DIP as it results from gender differences on mathematics achievement items. The data from a national administration of the Mathematics Usage subtest of the ACT Assessment was used in the investigation. The results indicated that there was the large instructional effect as predicted. However, there was also a smaller, gender effect on the performance of some items. (Author/DWH)

Interpretation of Differential Item Performance Accompanied by Gender Differences in Academic Background

Allen E. Doolittle

The American College Testing Program
P. O. Box 168
IOWA CITY, IOWA 52243

Paper presented at the annual meeting of the American Educational Research Association, New Orleans

April, 1984

ABSTRACT

The meaning of differential item performance (DIP), often referred to as item bias, is discussed. DIP is suggested here to encompass both item bias and instructional bias. Item bias is described as item invalidity that is unfair to certain population subgroups. Instructional bias is described as a valid reflection of group differences in instruction or background. Using data from a national administration of the ACT Assessment, this study investigated the plausibility of an instructional bias interpretation of DIP as it results from gender differences on mathematics achievement items. The results indicated that there was the large instructional effect as predicted, but there was also a smaller, gender effect on the performance of some items.

INTRODUCTION

During the past decade, there has been a great deal of work with the construct frequently referred to as "item bias". Most researchers now conclude that the term "item bias" is not sufficiently descriptive. Moreover, the common use of item bias as a synonym for terms such as differential item performance or item-group interaction is imprecise and can lead to a misunderstanding about the nature of the construct. Bias and, hence, item bias are value-laden terms which imply unfairness. In achievement tests, the construct can and frequently does exist without unfairness.

The confusion could be reduced by thinking of differential item performance (DIP) as a comprehensive term. In this sense, DIP refers to a kind of systematic item effect that works to the detriment of one group when compared to another.

Within the scope of this definition, it is possible for DIP to represent a systematic effect that is basically unfair, or actually biased against a group of examinees. In such an instance, differential item performance would be a form of item invalidity for that population subgroup and could be appropriately referred to as item bias. This situation could exist with an item that measures, in part, something unrelated to an intended test objective. In general, this is just poor measurement, but when groups differ in the knowledge measured by the item, it is also unfair to the deficient individuals or groups. The inappropriate inclusion in a test of an item measuring some characteristic not relevant to the test objectives is unfair or biased against those without the requisite background.

On the other hand, it is also possible for differential item performance to reflect group differences in the achievement of a relevant test objective. Here, DIP would again represent a systematic effect, but this time the difference in group performance would be a legitimate indication of group differences in instruction or preparation. For instance, if a test is a measure of general chemistry achievement, organic chemistry items would probably exhibit "bias" against equally able students with only an inorganic chemistry background. However, this is not bias in the sense of item unfairness. It is a valid reflection of insufficient instruction in organic chemistry. This form of DIP might be called "instructional bias".

Research has shown that male high school students as a group perform better than female high school students on mathematics achievement items (Fennema & Sherman, 1977). A plausible explanation is that male students typically receive more and a higher level of instruction in mathematics than do females. If so, one would expect that instances of differential item performance, in the form of instructional bias against females, might exist in mathematics achievement tests. Similar to the chemistry example cited earlier, instructional bias might be shown to exist for a complex algebra item if one group of students has been instructed in advanced algebra and another group of students, equal in general ability, has not.

The primary objective of this research was to investigate the nature of DIP as it relates to gender differences in mathematics achievement items. Since female students as a group tend to be less well prepared in mathematics than males, some instances of the instructional bias type of DIP were expected within a mathematics achievement test. In addition, it was believed that if instructional differences were minimized, little evidence of DIP would be found. A decided reduction in evidence of differential item performance from an uncontrolled situation to an instructionally controlled situation would suggest an instructional bias interpretation of the DIP found in the first analysis.

A secondary objective of this research was to present a multiple analysis approach to the study of DIP. It is unlikely that typical classification variables such as race and gender, in themselves, are major contributors to DIP. The multiple analysis approach used here may be useful for studying the extent to which DIP is a function of one of the usual classification variables or of some other variable more directly linked to item performance.

METHODOLOGY

Data Source

The data source for this research was the October, 1981 administration of the ACT Assessment Mathematics Usage test (ACTM) to a sample of 4,000 college-bound, high school students. Of these 4,000 students, 1,668 (41.7%) were male and 2,332 (58.3%) were female. As shown in Table 1, the mean ACTM scaled score for the males (16.1) was about one third of a standard deviation higher than the mean for the females (13.7). Also, the males averaged more semesters of mathematics coursework in a four-year high school career (6.2) than did females (5.6).

TABLE 1

Subgroup Descriptive Statistics

	Males (N=1668)		Females (N=2332)	
	ACTM	Math Sems	ACTM	Math Sems
Mean	16.1	6.2	13.7	5.6
S.D.	8.1	1.9	7.6	2.0

The Instrument

The ACT Assessment examination is an educational achievement test containing four subtests, one of which is Mathematics Usage (ACTM). The ACTM is a 40-item measure of mathematical reasoning ability. It emphasizes the solution of practical quantitative problems that are encountered in many postsecondary curricula and includes a sampling of mathematical techniques covered in high school courses. The test emphasizes quantitative reasoning rather than memoriza-

tion of formulas, knowledge of techniques, or computational skill.

Index of Differential Item Performance

A measure suggested by Linn and Harnisch (1981) was used as the index of DIP in this research. Although this measure is based on item response theory, it may be viewed as a "small sample" alternative to some of the more frequently studied IRT indices. To calculate the index, the item and ability parameters of the three-parameter logistic model are estimated for the total sample. The target group is then separated from the rest of the sample. (The target group can be any group of interest, but it is frequently thought of as a low-scoring group or one that may be adversely affected by DIP.) The difference is taken between each target group examinee's computed probability of correctly answering the item and the examinee's actual response to the item (1=correct; 0=incorrect). The index is this difference, standardized and averaged over all members of the target group. This index is considered a signed index. That is, the direction of the DIP is indicated. As calculated here, negative values represent DIP against the target group and positive values represent DIP favoring the target group.

There are a couple of potential advantages of this procedure over other IRT approaches. The primary advantage is its applicability to relatively small samples. The usual, 3-parameter IRT estimation procedures do not work well for samples of less than 1,000 (Linn & Harnisch, 1981; Wood & Lord, 1976). Since it is not uncommon for a subgroup to be this small or smaller, even when the overall size of the data set is quite large, the potential value of a small-sample, IRT alternative is clear. Another advantage, according to Linn and Harnisch, is that the index is weighted by the actual distribution of examinee ability estimates (thetas) in the target group. This is an advantage over procedures based on simple comparisons of item characteristic curves for specified groups which can suggest differences in situations where there may be few observations.[1] Previous research has shown the Linn and Harnisch measure to be a stable index and to be substantially correlated with other, perhaps more common, measures of DIP (Doolittle, 1983).

---

[1] There are other IRT indices that also share this particular advantage (see Shepard, Camilli, & Williams, 1983).

Instructional Background Indicator

Since a precise measure of instructional background in mathematics was not available, the members of the sample were classified on the basis of the number of semesters of mathematics instruction they received while in high school. For this research, those who reported at least six semesters of mathematics (in an eight-semester high school career) were considered the high background group; and those with less than six semesters were considered the low background group. As a result, 69.3% of the males and 57.9% of the females in the sample were placed in the high background category.

Research Design

This study consisted of several sub-analyses based on various divisions of the original sample (Table 2). As the sample was divided differently, the variation in the numbers of items with significant values of the index was expected to provide evidence as to the nature of the DIP. For an instructional bias interpretation, the instances of significant DIP were expected to fluctuate as follows (also Table 2, col. 1).

1. When the sample was divided on the basis of gender by itself, a moderate number of items was expected to show DIP.

2. When the sample was divided solely on the basis of instructional background (6-8 math. sems. vs. 0-5 math. sems.), a relatively large number of items were expected to exhibit DIP.

3. When the sample was divided on the basis of gender, but with instructional background held relatively constant for each group at either a high (a) or a low level (b), few of the items were expected to show DIP.

4. When the sample was divided on the basis of level of instruction, but with gender controlled, e.g. for males (a) and for females (b), relatively large numbers of items were expected to show DIP.

Column 2 of Table 2 suggests a possible outcome of these analyses if DIP, instead, were to reflect actual "item bias" against females.

Since the exact distribution of the Linn and Harnisch index is not known under the assumption of no DIP, an approxi-

mation to the distribution was calculated for each analysis. This procedure, suggested by Linn, Levine, Hastings, and Wardrop (1981), involved dividing a particular sample into essentially random halves and calculating the index on one of the halves as a pseudo target group. This was expected to represent a distribution of index values for the null hypothesis situation. The highest absolute value in this distribution was taken as the critical value. Since the various analyses in this study involved different subsamples, the approximate null hypothesis distribution was uniquely determined for each analysis.

TABLE 2

Hypothesized Numbers of Items Exhibiting DIF

Groups	Instruct. bias interpretation	Item bias interpretation
1. Male/female	moderate (baseline)	moderate (baseline)
2. Strong background/weak background	rel. high	low
3. a. Male, strong background/ female, strong background	low	moderate
b. Male, weak background/ female, weak background	low	moderate
4. a. Strong background, male/ weak background, male	rel. high	low
b. Strong background, female/ weak background, female	rel. high	low

RESULTS

The results shown in Table 3 were not entirely as expected. Three items were identified as biased in the male-female analysis (Analysis 1) and seven were so identified in the high instruction-low instruction analysis (2). In terms of numbers of biased items, these results were consistent with expectations. Similarly, the two items identified for the male low-female low analysis (3b), the seven items identified for the male high-female high analysis (4a), and the nine items identified for the female high-female low analysis (4b) were in line with expectations. Thus, five of the six analyses produced numbers of biased items basically as hypothesized. However, the male, high instruction-female, high instruction analysis (3a) produced seven biased items, substantially more items than was expected.

A further surprise was that the gender-oriented analyses (1,3a,3b) and the level-oriented analyses (2,4a,4b) did not yield similar results.[2] If gender were acting simply as a correlate for instruction, the signs for identified items should be the same across all analyses. However, without exception, items with significantly negative values of the index, for any of the gender-oriented analyses, had positive signs if also identified by any of the instruction-oriented analyses, and vice versa.

When the items with significant DIP were examined, some interesting patterns emerged. Relatively abstract items (strictly numbers and symbols), such as items 6 and 10, tended to favor the high instruction groups, and more concrete, arithmetic reasoning items (word problems), such as items 12,25,27 and 28, tended to favor the low instruction groups. This relationship seems intuitively plausible since the most instruction in abstract mathematics is likely to be received by the more advanced students -- those with sufficient prerequisite coursework. Thus, it would seem likely

---

[2] The six separate analyses may be logically grouped based on their focus. The gender-oriented analyses are those that compare the item performance of males and females regardless of whether or not level of instruction has been controlled. Similarly, the instructional level analyses are those that compare the item performance of high and low instructional level examinees regardless of whether or not gender has been held constant.

that high instruction students would do relatively well on these items. Conversely, it would seem to follow that low instruction examinees would perform relatively better on the more concrete mathematics items.

Although the indications of DIP were less strong for the gender-oriented analyses than for the instructional level analyses, again there seemed to be some notable patterns. Geometry items (items 3 and 11) and, to some extent, arithmetic reasoning problems (items 12 and 25) seemed to adversely affect the performance of females. Why these types of items appeared relatively more difficult for females, however, is not readily apparent.

TABLE 3

Significant DIP in ACTM Items*

	Analysis					
	1	2	3a	3b	4a	4b
Item	M/F(.06)	H/L(.06)	MH/FH(.06)	ML/FL(.06)	MH/ML(.08)	FH/FL(.05)
3	-.07		-.10			
4						.06
6		-.19			-.22	-.16
7		-.10	.07		-.10	-.11
10		-.15	.08		-.14	-.15
11	-.07		-.08	-.07		
12	-.08	.07	-.07	-.10	.16	
13			.07			-.05
14						.06
16		-.07			-.12	
25			-.07			.06
27		.08			.15	
28		.09			.10	.09
37						.07
# Sig.:	3	7	7	2	7	9
Predicted:	mod	high	low	low	high	high

* Significance determined by comparison to the largest absolute value of the statistic calculated in a random, null hypothesis situation. The obtained critical value of the index is shown in parenthesis for each analysis.

The analysis headings are:

M - male	F - female
H - high instruction	L - low instruction
MH - male, high instruct.	FH - female, high instruct.
ML - male, low instruct,	FL - female, low instruct.

The second group under each heading is the target group. A negative value represents DIP to the disadvantage of the target group while a positive value is DIP to the relative benefit of the target group.

DISCUSSION

Although the primary intention of this research was to clarify the nature of gender-related DIP in mathematics achievement items, the results were not quite as expected. Evidence of DIP was found in the analyses focused on instructional level and, to a slightly lesser extent, in the gender-oriented analyses. However, the fact that the direction of the DIP was not consistent for the same items across analyses suggests that, in this situation, gender was more than a simple correlate of instructional level. The notion that gender-related DIP among mathematics achievement items is merely due to gender differences in instructional level was not supported.

The evident instances of DIP, seemingly due to gender and not instructional level, suggest the possibility of at least two alternative hypotheses. Perhaps the measure of instructional level used in the study was inadequate. That is, quantity or number of high school mathematics courses may not have been an appropriate measure. If there is a substantial group to group discrepancy in the type or quality of instruction received, a measure based on quantity would indeed seem to be inadequate. Possibly this could have been enough of a problem to substantially impact these results.

Another possible explanation is that there may be items that do, in fact, perform differently for males and females, regardless of instructional level. This may be akin to true item bias but, more likely, it suggests the existence of some other background variable, like examinee expectations, acculturation, or motivation, that could differentially affect group performance.

The fact that there seemed to be certain groups of items that favored one group over another indicates that future research emphasizing item type or item content might be fruitful. A possible avenue for this kind of research is the experimental design approach suggested by Schmeiser (1983). When there are indications that certain types of items may be biased against one group or another, the use of relevant analysis of variance procedures could provide a practical means for examining the differential impact of, for example, geometry items on gender groups. From an exploratory research perspective, it might also be helpful to simply ask female examinees why they had trouble with certain geometry and arithmetic reasoning items.

The secondary objective of this study was to present a multiple analysis approach for the investigation of DIP. Such an approach seems particularly useful for learning more about the nature of DIP. Many "item bias" studies have stopped at the level of the male-female analysis in this study. That is, if any evidence of DIP is found, there is a tendency to conclude that particular items are biased against one group or another. It is argued here that such an interpretation may be premature. Although the results of this research did not clearly demonstrate the potential problems with the typical item bias study, the suggested methodology should provide a means to explore DIP at a deeper level. The multiple analysis approach can be used to investigate likely causal variables, such as instruction, in addition to the more typical group classification variables, such as race or gender.

REFERENCES

Doolittle, A. E. The reliability of measuring differential item performance. ERIC #ED 234061. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1983.

Fennema, E. L., and Sherman, J. Sex-related differences in mathematics achievement, spatial visualization and affective factors. American Educational Research Journal. 1977, 14, 51-71.

Linn, R. L., Levine, M. V., Hastings, C. N., and Wardrop, J. L. Item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5(2), 159-173.

Linn, R. L., and Harnisch, D. L. Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 1981, 18(2), 109-118.

Schmeiser, C. B. Differences between black and white examinee performance on the ACT Assessment examination as a function of the racial orientation of test content. Doctoral dissertation, The University of Iowa, Iowa City, 1983.

Shepard, L., Camilli, G., and Williams, D. M. Accounting for statistical artifacts in item bias research. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1983.

Wood, R. L., and Lord, F. M. A User's Guide to LOGIST. Research Memorandum. Princeton, N. J.: Educational Testing Service, 1976.