DOCUMENT RESUME

ED 263 218 TM 850 714

AUTHOR Doolittle, Allen E.
TITLE Understanding Differential Item Performance as a Consequence of Gender Differences in Academic Background.
PUB DATE Apr 85
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; High Schools; Item Analysis; Mathematics Achievement; *Mathematics Instruction; *Sex Differences; *Spatial Ability; Statistical Analysis; *Test Bias; Testing Problems
IDENTIFIERS *ACT Assessment Mathematics Usage Test; *Differential Item Performance

ABSTRACT

Differential item performance (DIP) is discussed as a concept that does not necessarily imply item bias or unfairness to subgroups of examinees. With curriculum-based achievement tests, DIP is presented as a valid reflection of group differences in requisite skills and instruction. Using data from a national testing of the ACT Assessment, this study investigated the hypothesis that differences in high school instruction are the cause of gender-based DIP with mathematics achievement items. Research was conducted on a random sample of 2,669 college-bound, high school seniors from the October 1983 administration of the ACT Assessment Mathematics Usage Test (ACTM). Forty-five percent were males and fifty-five percent were females. The mean ACTM scaled score for the males was about one-half of a standard deviation higher than the mean for the females. The males averaged more semesters of mathematics coursework than did the females, and a higher percentage of males reported advanced or accelerated high school math courses. The results indicated that there was a substantial gender effect that could not be explained by instructional differences at the secondary school level. Geometry and word problems tended to have the greatest negative impact on female examinees. (Author/LMO)

Understanding Differential Item Performance as a Consequence of Gender Differences in Academic Background

Allen E. Doolittle

The American College Testing Program
P.O. Box 168
Iowa City, IA 52243

A paper presented at the annual meeting of the American Educational Research Association.

Chicago, April 1985

ABSTRACT

Differential item performance (DIP) is discussed as a concept that does not necessarily imply item bias or unfairness to subgroups of examinees. With curriculum-based achievement tests, DIP is presented as a valid reflection of group differences in requisite skills and instruction. Using data from a national testing of the ACT Assessment, this study investigated the hypothesis that differences in high school instruction are the cause of gender-based DIP with mathematics achievement items. The results indicated that there was a substantial gender effect that could not be explained by instructional differences at the secondary school level. Geometry and word problems tended to have the greatest negative impact on female examinees.

In recent years, there has been a great deal of work with the construct frequently referred to as "item bias". Many 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 item bias are value-laden terms that 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 items that are basically unfair, or actually biased against a group of examinees. On the other hand, it is also possible for DIP to fairly 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 preparation or instruction. For instance, in a test 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 would be simply the manifestation of an instrtctional effect.

Research has shown that male high school students as a group perform better than female high school students on mathematics achievement tests (Benbow and Stanley, 1980; Fennema and Carpenter, 1981). Benbow and Stanley (1980) suggest that these differences may be due to gender differences in spatial skills. Another possible explanation is that male students typically receive more and/or a higher level of instruction in mathematics than do females. If the latter were true, one would expect that instances of differential item performance, in the form of an instructional effect against females, might exist in mathematics achievement tests. As in the case of the chemistry example cited earlier, instructional bias might be shown to exist for a higher level mathematics item if one group of students has been appropriately instructed in mathematics and another group of students, equal in general ability, has not.

An earlier study (Doolittle, 1984), using data from one national administration of the ACT Assessment Mathematics Usage Test (ACTM), investigated the plausibility of an instructional interpretation of DIP in a situation where gender differences were known to exist. In the study, an index suggested by Linn and Harnisch (1981) was used to detect differentially performing items in six separate analyses. The analyses were based on comparisons of different subgroups defined by various combinations of gender and academic background taken from the original sample.

The results provided support for the seemingly self-evident notion that differences in instructional background have a strong influence on mathematics achievement. However, the results did not support the notion of gender-based DIP in mathematics achievement due primarily to differences in instructional background. As predicted, more items were found with significant DIP when the groups were defined by instructional differences than when they were defined by gender. But contrary to the hypothesis that gender was simply a surrogate for level of mathematics instruction was the fact that the direction of the DIP was often different for females than it was for the low instruction group. In other words, items that tended to work to the relative disadvantage of females were often found to disproportionately favor the low instruction groups, and vice versa.

The measure of instructional background used in the Doolittle (1984) study was the number of semesters of mathematics instruction received in high school. Those in the sample 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. A problem with this measure is that although it is perhaps a good measure of quantity of mathematics instruction, it says nothing about the type of instruction and is probably not a good indicator of quality of instruction. It seemed very possible that there could be substantial differences in the instructional backgrounds of students having the same number of mathematics courses to their credit. Perhaps a different measure of instructional level might have provided more clarity to the results of the study.

The nature of the data set may also have contributed to the unusual findings of that research. The research was conducted on a sample that included equal numbers of black and white students and was not representative of the ACT Assessment examinee population. Since females and "low-instruction" examinees were over-represented among blacks, a confounding of the results seemed possible. It was not clear what the outcome of the study would have been had the research been done on a more representative sample of students.

The primary objective of the present research was to continue the investigation of DIP as it relates to gender differences in mathematics achievement items, using a different definition of mathematics instructional level and a sample representative of the ACT examinee population. A secondary objective was to examine the types of items found to perform differently in different examinee groups for possible clues leading to a better understanding of DIP.

METHODOLOGY

Data Source

This research was conducted on an essentially random sample of 2,669 college-bound, high school seniors from the October 1983 administration of the ACT Assessment Mathematics Usage Test (ACTM) in the state of Ohio. Of these 2,669 students, 1,210 (45.3%) were male and 1,459 (54.7%) were female. As shown in Table 1, the mean ACTM scaled score for the males (23.3) was about one-half of a standard deviation higher than the mean for the females (19.2); and the males averaged more semesters of mathematics coursework in a four-year high school career (7.2) than did the females (6.6). Additionally, a higher percentage of the males reported advanced or accelerated high school math courses (37.1%) than did females (28.7%).

TABLE 1

Subgroup Descriptive Statistics

	Males (N=1210)		Females (N=1459)	
	ACTM	Math Sems	ACTM	Math Sems
Mean	23.3	7.2	19.2	6.6
S.D.	8.0	1.4	7.6	1.7

	High Males (N=915)		Low Males (N=295)		High Females (N=842)		Low Females (N=617)	
	ACTM	Math Sems	ACTM	Math Sems	ACTM	Math Sems	ACTM	Math Sems
Mean	25.4	7.8	17.3	5.3	22.3	7.7	15.2	5.2
S.D.	7.3	0.6	6.8	1.4	7.3	0.9	6.0	1.4

The Instrument

The ACT Assessment program contains educational achievement tests in four content areas, one of which is Mathematics Usage (ACTM). The ACTM is a 40-item, 50-minute 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 memorization of formulas, knowledge of techniques, or computational skill. In general, the mathematical skills required for the test involve proficiencies emphasized in high school plane geometry and first- and second-year algebra. Six types of items are included in the test and are described below.

1. **Arithmetic and Algebraic Operations (AAO).** The items in this category explicitly describe operations to be performed by the student. The operations include manipulating and simplifying expressions containing arithmetic or algebraic fractions, performing basic operations in polynomials, solving linear equations in one unknown, and performing operations on signed numbers.

2. **Arithmetic and Algebraic Reasoning (AAR).** These word problems present practical situations in which algebraic and/or arithmetic reasoning is required. The problems require the student to interpret the question and either to solve the problem or to find an approach to its solution.

3. **Geometry (G).** The items in this category cover such topics as measurement of lines and plane surfaces, properties of polygons, the Pythagorean theorem, and relationships involving circles. Both formal and applied problems are included.

4. **Intermediate Algebra (IA).** The items in this category cover such topics as dependence and variation of quantities related by specific formulas, arithmetic and geometric series, simultaneous equations, inequalities, exponents, radicals, graphs of equations, and quadratic equations.

5. **Number and Numeration Concepts (NNS).** The items in this category cover such topics as rational and irrational numbers,

set properties and operations, scientific notation, prime and composite numbers, numeration systems with bases other than 10, and absolute value.

6. **Advanced Topics (AT).** The items in this category cover such topics as trigonometric functions, permutations and combinations, probability, statistics, and logic. Only simple applications of the skills implied by these topics are tested.

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 the three-parameter logistic model (Birnbaum, 1968), it may be viewed as a "small sample" alternative to some of the more frequently studied item response theory (IRT) indices. Applicability to small samples is considered to be an advantage, since it is not uncommon for a subgroup to be small, even when the overall size of the data set is reasonably large.

Like other indices of DIP, the Linn and Harnisch index is only a relative, not an absolute, measure of DIP. That is, the index assumes that the total test score is an unbiased measure of ability or achievement. With this assumption, DIP exists when the performance of an item for a particular group is not in line with the overall performance of the group.

To calculate the Linn and Harnisch index, the item and ability parameters of the three-parameter item response theory model are estimated for the total sample. The target group is then separated from the rest of the sample. The difference is taken between each target group examinee's probability of correct response to an 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, sign indicates the direction of the DIP. As calculated here, negative values represent DIP against the target group and positive values represent DIP favoring the target group. Previous research has shown the Linn and Harnisch measure to be a reliable index and to be substantially correlated with other, perhaps more common, measures of DIP (Doolittle, 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 two self-reported variables: number of semesters of math instruction received in high school and participation in advanced or accelerated high school mathematics coursework. Students reporting either 8 semesters of high school math (out of a possible 8) or participation in accelerated math, or both, were categorized as having a high level of mathematics background. Those who did not meet either of these criteria were, for the purposes of this study, considered the low background group. Consequently, 75.6% of the males and 57.7% of the females were placed in the high background category. Table 1 depicts mean score on ACTM as well as mean math semesters for each background and gender category. About 50% of both the males and the females in the high background groups had been involved in accelerated high school math courses.

Research Design

This study was done in two stages. The first consisted of the three analyses described below, with a focus on detection of substantial DIP.

1. Background level. The data were analyzed based on differences in level of mathematics instruction. Males and females were grouped together and then dichotomously categorized according to background. Items were identified that tended to relatively favor one or the other of the two groups.

2. Gender. Students were separated based on gender alone, each group consisting of some students with high and some with low background levels. Items were identified that tended to relatively benefit either of the groups.

3. High background and gender. Only students classified as "high background" were included in this analysis. The data were then analyzed with gender as the primary variable, essentially controlling for instructional level. Items with substantial DIP were again identified.

Since the exact distribution of the Linn and Harnisch index is not known under the assumption of no DIP, an approximation to the distribution was calculated for each analysis. The utilized procedure was a modification of a procedure suggested by Linn, Levine, Hastings, and Wardrop (1981) that involved dividing each sample into essentially random halves and calculating the DIP 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 mean and standard deviation of the obtained index values were used in conjunction with normal distribution tables to determine approximate critical values of the statistic for $\alpha = .01$. Since each analysis involved different subsamples, the approximate null hypothesis distribution was uniquely determined for each case.

The second stage of the study involved a review of the types of items identified with DIP in the previous analyses. For example the Doolittle (1984) study reported that AAO items (primarily numerical operations) tended to benefit high instruction students and that AAR items (word problems) tended to favor low instruction students. Additionally, AAR and Geometry items tended to show DIP that favored male students.

Results

The results shown in Table 2 indicate that 16 of the 40 items were identified with substantial DIP in the background level analysis and 12 items were identified in the gender analysis. Without exception, where items were identified with DIP for both analyses, the direction of the DIP was not the same. In other words, an item, such as item 2, identified as "biased" in favor of the low instruction group, was biased against females. In fact, the product moment correlation between the two sets of indices in analyses 1 and 2, over all 40 items, was -0.54 ($p < .001$). This result runs counter to the notion that gender-related DIP is simply a reflection of differences in instruction, but it is consistent with the Doolittle (1984) study.

When the gender analysis was repeated, after controlling background at the high level, results similar to the uncontrolled, overall gender analysis were obtained. Eleven of the 40 items indicated significant levels of DIP. Seven of these 11 were also identified in the overall gender analysis. The four items that were identified in the controlled analysis, but not in the overall analysis, were all found at the end of the test and were some of the most difficult items on the test. Perhaps the fact that DIP showed up there might be due to the ability of the high background students, as opposed to the total sample, being more suitable for the proper functioning of the harder items.

When the items with significant DIP were examined, some interesting patterns, shown in Table 3, became apparent. AAR items (word problems), tended to favor the low instruction group and the more abstract, Intermediate Algebra (IA) items, such as 10 and 19, tended to favor the high background group. This relationship seems intuitively plausible since the most instruction in relatively abstract mathematical concepts is likely to be received by the more advanced students. Thus it seems reasonable to expect that the high background students would do relatively well on these items. Conversely, it would seem to follow that the low background examinees would perform relatively better on word problems that are perhaps less dependent upon advanced instruction. Similar results were obtained in the earlier Doolittle study.

Different, but no less convincing, patterns were found for the gender analysis. Geometry items, such as items 24, 28, and 31, and AAR (word problem) items seemed to favor males, while the remaining item categories tended to favor females. These patterns are consistent with the results of Doolittle (1984), Pattison and Grieve (1984), and Smith (1984).

Table 2

Significant DIP in ACTM Items*

		Analysis		
		1	2	3
		(BACKGROUND)	(GENDER)	
Item	Classification	H/L (+.06)	M/F (+.05)	MH/FH (+.08)
2	AAR	.07	-.09	-.10
4	NNS		.09	
5	AAR	-.06	.07	.08
6	AAR	.10		
8	AAR	.09		
10	IA	-.15	.07	.08
11	AAR	.07	-.09	-.10
12	IA	.10		
13	AAR		-.07	
15	AAR		-.06	
16	AAO	-.09		
17	AAR	.10		
19	IA	-.12		
24	G		-.06	
25	NNS	-.07		
26	AAR	.06		
28	G		-.07	-.09
29	AT	-.13	.09	.13
30	IA	-.07	.06	
31	G		-.12	-.16
32	AAR	.09		
35	G	.06		
36	IA			-.09
37	G			-.08
38	G			-.09
39	NNS			-.10

* Significance determined by comparison to the distribution of the index calculated in a random, null hypothesis situation. The critical values for a two-tailed test of the index (α approximately .01) are shown in parentheses 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.

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

Table 3

Differential Item Performance by Item Category

	Level Analysis			Gender Analysis		
	(Group favored)			(Group favored)		
Item Category	High	None	Low	Male	None	Female
AAO	1	3	0	0	4	0
AAR	1	6	7	4	9	1
IA	3	4	1	0	6	2
G	0	7	1	3	5	0
NNS	1	3	0	0	3	1
AT	1	1	0	0	1	1

DISCUSSION

It is apparent from this study that gender-based, differential item performance in mathematics is not a simple consequence of group differences in mathematics background. If it were, the direction of the DIP would be expected to be similar for low background examinees and females (since they tend to have weaker math backgrounds than males). Since the direction of the DIP was not similar for females and low background students, it appears that some items may show "gender bias" despite the confounding influence of a gender by background interaction. The results of this study indicate that, with respect to overall mathematics achievement, certain kinds of items are relatively easier for males while others are relatively easier for females.

Since this study does not support the notion of gender-based DIP as a function of high school math background, the cause or causes of gender differences in performance are not clear. However, it does not seem particularly useful to think of tests like the ACT Assessment as "biased" in the sense of unfairness. The tests are carefully assembled from specific objectives that are tied directly to high school mathematics curricula. It seems likely that differences in the learning of mathematics concepts do exist among high school seniors and that these differences are being accurately assessed by items in the ACT Mathematics Usage Test.

Why such differences occur between males and females may be the subject of additional speculation and research. Perhaps gender differences in acculturation and instruction are well established prior to the high school years; or, it may be that there exist specific differences in spatial skills (Benbow and Stanley, 1980; Pattison and Grieve, 1984) that may affect performance on geometry and some other mathematics items. Further research in the assessment of gender differences in mathematics items at the secondary level might focus on variables beyond the rough classification of items such as degree of spatial content, verbal content, and abstractness. Perhaps this kind of research could yield a better understanding of existing differential item performance.

REFERENCES

Benbow, C. P., and Stanley, T. P. Sex differences in mathematical ability: fact or artifact? Science, 1980, 210, 1262-1264.

Birnbaum, A. Some latent trait models and their use in inferring an examinees ability. In Statistical Theories of Mental Test Scores by Lord, F. M. and Novick, M. R., 1968, pp. 404-405.

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.

Doolittle, A. E. Interpretation of differential item performance accompanied by gender differences in academic background. ERIC #ED 247237. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1984.

Fennema, E., and Carpenter, T. P. Sex-related differences in mathematics: results from the National Assessment. Mathematics Teacher, 1981, 74, 554-559.

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.

Pattison, P. and Grieve, N. Do spatial skills contribute to sex differences in different types of mathematical problems? Journal of Educational Psychology, 1984, 76(4), 678-689.