DOCUMENT RESUME

ED 067 404 TM 001 799

AUTHOR Echternacht, Gary
TITLE A Quick Method for Determining Test Bias.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO RB-72-17
PUB DATE Apr 72
NOTE 16p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Evaluation Criteria; Group Tests; *Probability Theory; *Research Methodology; *Statistical Analysis; *Test Bias; *Test Construction

ABSTRACT

The problem of test bias has been a growing concern in recent years. Of the several available methods for determining test bias, probably the most effective means involves collecting criterion information. This data collection process often provides a considerable barrier to the researcher, especially for the small test user and for someone who needs an immediate solution to a test bias question. This paper presents a method for identifying and analyzing the nature of test bias. This method is intended as only a preliminary analysis prior to, or concurrently to, a criterion data collection process. (Author)

RESEARCH BULLETIN

RB-72-17

A QUICK METHOD FOR DETERMINING TEST BIAS

Gary Echternacht

Educational Testing Service
Princeton, New Jersey

April 1972

A QUICK METHOD FOR DETERMINING TEST BIAS

Gary Echternacht

Abstract

The problem of test bias has been a growing concern in recent years. Of the several available methods for determining test bias, probably the most effective means involves collecting criterion information. This data collection process often provides a considerable barrier to the researcher, especially for the small test user and for someone who needs an immediate solution to a test bias question. This paper presents a method for identifying and analyzing the nature of test bias. This method is intended as only a preliminary analysis prior to, or concurrently to, a criterion data collection process.

A QUICK METHOD FOR DETERMINING TEST BIAS

Gary Echternacht

In the recent past, there has been a growing concern about the "fairness" of psychological tests with respect to various groups in the testing population. Usually, these concerns can be classified under the general notion of test bias. The issue of test bias will most likely become even more important in the future as test users are required to show that their tests are not biased against any subgroup of the testing population.

The past literature has seen a number of studies of test bias. Most notable among those are studies by Cardall and Coffman (1964) and Cleary and Hilton (1968). A rather comprehensive statistical review of the problem of test bias has been given by Potthoff (1966), where he discusses the problem and presents operational definitions in two different research settings: in the absence and presence of a criterion. Most researchers will generally agree that the definition of test bias and the methodology for determining its extent is most powerful when a criterion is present. The usual definition, in this case, says that there is no test bias if individuals from different groups who have the same test scores also have the same expected criterion scores.

Although the above definition provides the researcher with considerable problems that will not be mentioned here (Darlington, 1971; Potthoff, 1966; Thorndike, 1971), in actual practice two difficulties stand out in attacking a test bias question with a criterion: (1) it is often difficult to conceptualize an adequate criterion variable or variables, especially when one is using the test to predict something akin to academic "success," and (2) criterion data are often difficult and expensive to obtain, in terms of both time and

money. It, therefore, seems reasonable to assume that initial attempts to study test bias will, most likely, involve a model not utilizing a criterion variable.

In the absence of a criterion variable, many researchers get stalled in defining test bias. Although there seems to be no generally suitable way to define test bias in the absence of a criterion, the concept of item-group interaction appears to be closely related. This allows for groups to differ, but requires that difference be constant for each item in the test. Many researchers conjure up a notion of item-group interaction as being that effect tested in a repeated-measures analysis of variance, where items are the repeated measures and the groups classified in some factorial structure. The significance tests resulting from this type of analysis are somewhat suspect, in that the observations are nonnormal discrete random variables and the cell variances tend to be nonhomogeneous.

There are many problems in relating item-group interaction to test bias. The main difficulty seems to be that there is no generally suitable way of defining interaction. In some cases, what is concluded to be item-group interaction is merely balancing where the test constructor includes items relatively more favorable to one group in order to make the total scores closer for the groups under consideration. The test constructor must decide the extent of balancing present in any test. In other cases, where no item-group interaction is concluded, test bias may be uniformly present in each item. Also, item-group interaction deals with group differences, but differences by what measure? Thus, the reader should see that an examination of test bias without a criterion can result in only very tentative results, results that provide more of a cue about the items involved in a possible bias, rather than the actual existence or nonexistence of test bias.

Potthoff (1966) proposed a conservative multivariate technique, that seemed most promising, in assessing item-group interaction. Basically, Potthoff's method involved taking a sample and calculating a significance test for each of the $\binom{g}{2}\binom{k}{2}$ item-group pairs, where g and k indicate the number of groups and items respectively. An arbitrary number, M, of the item-group pairs with the lowest significance levels are chosen ($5 \leq M \leq 50$). An independent sample is drawn, and significance levels are calculated for these M items. A significance level for all of the item-group pairs, tested simultaneously, is $M\alpha_2$ where α_2 is the lowest of the M significance levels. The difficulty in this procedure is that the computational power of completing such a test is not available for many users, especially if either g or k is large, and it is expensive to set up. For that reason, this paper proposes a computationally easy method for determining the nature and extent of test bias in a test. The general notion of item-group interaction is used in the framework of the absence of a criterion variable. It should be noted that this technique is designed to serve as a prelude to a more penetrating study of test bias, usually one where criterion information is present. If the notion of item-group interaction seems to be synonymous with that of test bias, then the procedure is quite reasonable.

Methodology

In the general setup, we consider g groups of people identifiable by some characteristic or combination of characteristics, i.e., sex, national origin, race, socioeconomic status, etc. Consider further a test of m items. The data consist of all p-values, p_{ij}, $i = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, g$, which represents the proportion of people in the jth group answering the ith item correctly. We would like to say that there is no item-group interaction

(test bias) if the differences, $p_{ij} - p'_{ij}$, are constant over the i items in the population. Potthoff (1966) has shown that some quirks can occur if p-values are used which might cause one to conclude the presence of an interaction when such an interaction does not exist.

As an example of the quirk in the definition, consider the case of two groups, where a constant difference in p-values for all items is required for a null hypothesis of no item-group interaction. Suppose, further, that the difference $p_{2k} - p_{1k} = 0.2$ for all k items. If we choose to add an item whose difficulty was .15 for Group 2, this would require us to conclude the existence of an interaction since it is impossible to have a p-value of -.05 for Group 1. On the other hand, if we introduce an item that has a difficulty of .90 for Group 1, then again interaction must be concluded since a difficulty of 1.10 would be required for Group 2. If the difficulties are sufficiently far from 0 and 1, these quirks disappear. On the other hand, most test constructors place both a few very easy items and a few very difficult items in almost every test so that the quirk previously mentioned occurs very frequently. For that reason, the sample p_{ij}'s are transformed to a quantity termed delta (Conrad, 1948), denoted by Δ_{ij}, using the transformation

$$\Delta_{ij} = 13 - 4z_{p_{ij}} ,$$

where $z_{p_{ij}}$ is such that

$$p_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{p_{ij}}} e^{-z^2/2} dz ;$$

the value of z corresponding to a cumulative normal ordinate value of p_{ij}. Next, we restate the previous definition of item-group interaction to say that there is no item-group interaction (test bias) if the differences $\Delta_{ij} - \Delta'_{ij}$ are constant for each i, $i = 1, 2, \ldots, m$, and depend only on j and j'. The Δ_{ij} transformation (delta transformation) is a transformation not uncommonly found in various item analysis packages and is usually rounded to only one decimal. It is also referred to as the item difficulty index.

Under a null hypothesis of no item-group interaction, the sample differences $\Delta_{ij} - \Delta'_{ij}$ should be distributed normally with some unknown mean and variance for each group j and j'. If evidence can be gathered to the effect that this is not the case, the null hypothesis can be rejected.

In order to determine whether the differences in item deltas are normally distributed, two steps should be undertaken. First, item delta differences should be calculated for certain pairs of groups. The nature of the pairing will be discussed later. Once all these differences have been calculated, these differences are plotted on normal probability paper for each pair of groups. In order to obtain points to plot, first one obtains the order statistics, $y_{(s)}$, for each pair of groups. The order statistics, $y_{(s)}$, are then paired with the values $s/(m + 1)$ for the complete specification of the point. Ties are handled by considering j to be the number of differences less than j itself. Thus, differences are ranked from lowest to highest, where s differences share the same rank r, the following observation has rank $r + s$. A good elementary description of this plotting can be found in Schmitt (1969).

If the differences truly follow a normal distribution, the plots will be in a straight line. In order to determine whether the plots fall on a straight line when deltas are plotted on normal probability paper, the method suggested by Lilliefors (1967) is appropriate. Basically, this is a modification of the Kolmogorov-Smirnov test for normality where the hypothesized mean and variance are calculated from the sample data. The mean of the differences in deltas is calculated as is the standard deviation. The mean is plotted at 50 and the mean plus one standard deviation at 84. A straight line is drawn connecting the two points, which represents the hypothetical cumulative normal distribution. A band is drawn around the line in accordance with the number of items and the desired significance level. If any points fall outside the band, normality is rejected, and item-group interaction (test bias) is concluded.

If just two groups are being considered, only one set of plots need be constructed. If more than two groups are considered, there is a problem of which group pairs to consider. The problem arises from the fact that only $g - 1$ of the group pairs are independent. For example, if we consider four groups, any set of item differences can be constructed by considering only the differences between Group 1 and Group 2, Group 1 and Group 3, and Group 1 and Group 4. The general problem consists of choosing which group differences to plot.

Quite frequently, the groups can be classified into a factorial structure. For example, one might consider four groups defined by all combinations of sex and race (consider only Caucasian and Negroid here). In this case, one possibility might be to examine racial bias by: (1) plotting differences in delta values for Caucasian and Negroid Males, (2) plotting differences

in delta values for Caucasian and Negroid females, and (3) pooling the data over race and plotting differences for males and females. The first two plots are tests of racial item-group interaction (test bias) within sex, while the third is a test of the same between sexes.

The above setup is rather arbitrary and depends only on the researcher. One might, equally well, consider item-group interaction within race and between races. It should be noted that once the first two comparisons have been established, the third follows automatically.

Example

An example is given here for a subtest from a large national testing program. The section contains 30 items of the reading comprehension item type. In this example, the primary interest was in identifying racial bias. There were four groups under study, defined by sex and race. Those groups were white females, black males, white males, and black females. Deltas were calculated for each item in each group as part of an item analysis. White male vs. black male, white female vs. black female, and male vs. female (the only possible independent comparison remaining) combinations were formed and differences in deltas calculated for each pairing on each item. These differences appear in Table 1 along with their ordering and the values of

---

Insert Table 1 about here

---

100 x s/(m + 1). Plots on 8 1/2 x 11 normal probability paper are presented in Figures 1-3. The solid line represents the hypothetical normal with

---

Insert Figures 1-3 about here

---

the sample mean and variance as the parameters. The dashed lines represent the critical values for a sample size of 30 at the .05 level of significance. In this case the critical value is .161, or 1.61 using the scale indicated on the figures. If any point falls outside the band formed by the dashed lines, test bias is concluded. This occurs in each of the three figures, with the point (6, -19) of Figure 1 differing from the hypothesized line by exactly 1.6, the critical value with only two significant digits.

One is led to the conclusion that this test is biased, assuming item-group interaction is an acceptable definition of test bias. This bias is represented both as racial bias within sex and between sexes. The nature of the bias cannot actually be specified, as the hypothesis tested here was one of existence rather than nature. A clue though might be that points falling below the bands are indicators of bias against the lower scoring group (group with the highest delta) and points falling above the bands of indicators of bias against the higher scoring group (group with the lowest delta). Using this characterization, a test is most likely not to be biased against any one particular group, but rather biased in a complex manner. Bias could occur against both groups as in Figures 2 and 3.

Summary

This paper presents a quick method for assessing test bias. It assumes no criterion is present and equates the notion of item-group interaction with that of test bias. The procedure is to:

1. Transform the item p -values to delta values.
2. Form independent pairs of groups and obtain delta differences.
3. Plot these differences on normal probability paper.

4. Obtain the mean and variance of the differences and plot a hypothetical normal distribution with the obtained mean and variance as parameters.

5. Draw bands around the hypothetical line whose width is determined by the sample size and significance level, and conclude test bias if any points fall outside the bands.

If the notion of item-group interaction is accepted as equivalent to test bias, this procedure represents a valid test of significance for the existence of test bias. The visual display further provides clues as to the nature of the bias if such is present.

References

Cardall, C., & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. College Board Research and Development Reports 64-5, No. 9 and ETS Research Bulletin 64-63. Princeton, N. J.: Educational Testing Service, 1964.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 71-75.

Conrad, H. S. Characteristics and uses of ten analysis data. Psychological Monographs, 1948, 62 (Whole No. 295).

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Lilliefors, H. W. The Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 1967, 62, 399-402.

Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests. Institute of Statistics Mimeo Series No. 479. Chapel Hill: University of North Carolina, Department of Statistics, 1966.

Schmitt, S. A. Measuring uncertainty: An elementary introduction to Bayesian statistics. Reading, Mass.: Addison-Wesley, 1969.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Table 1

Delta Differences and Plotting Points

Item No.	White Males–Black Males			White Females–Black Females			Males–Females		
	Difference x 10	Rank	100 $(\frac{s}{m+1})$	Difference x 10	Rank	100 $(\frac{s}{m+1})$	Difference x 10	Rank	100 $(\frac{s}{m+1})$
1	-21	2	6	-34	3	10	- 1	27	87
2	-18	7	23	-22	10	32	-10	11	35
3	-21	2	6	-26	8	26	-18	1	3
4	- 8	23	74	-20	12	39	-11	7	23
5	-21	2	6	-17	18	58	-12	4	13
6	-18	7	23	-18	15	48	-11	7	23
7	-20	6	19	-27	7	23	- 9	16	52
8	-17	9	29	-32	5	16	-10	11	35
9	- 8	23	74	-18	15	48	- 1	27	87
10	-26	1	3	-36	1	3	-10	11	35
11	-17	9	29	-35	2	6	-10	11	35
12	-13	16	52	-28	6	19	-18	1	3
13	-15	12	39	-19	14	45	- 9	16	52
14	-16	11	35	-16	19	61	-17	3	10
15	-15	12	39	-15	21	68	- 6	20	65
16	-12	17	55	-16	19	61	-11	7	23
17	-14	15	48	-18	15	48	- 6	20	65
18	-13	16	52	- 9	27	87	- 1	27	87
19	- 4	28	90	-22	10	32	- 7	19	61
20	- 2	30	97	4	30	97	- 6	20	65
21	-13	16	52	-20	12	39	- 9	16	52
22	- 3	29	94	-10	26	84	-12	4	13
23	- 7	27	87	-14	23	74	-10	11	35
24	-21	2	6	-34	3	10	- 4	24	77
25	-10	18	58	-25	9	29	-12	4	13
26	- 8	23	74	- 8	29	94	- 4	24	77
27	- 9	22	71	- 9	27	87	- 1	27	87
28	-10	18	58	-13	24	77	- 5	23	74
29	-15	12	39	-11	25	81	- 3	26	84
30	- 8	23	74	-15	21	68	-11	7	23

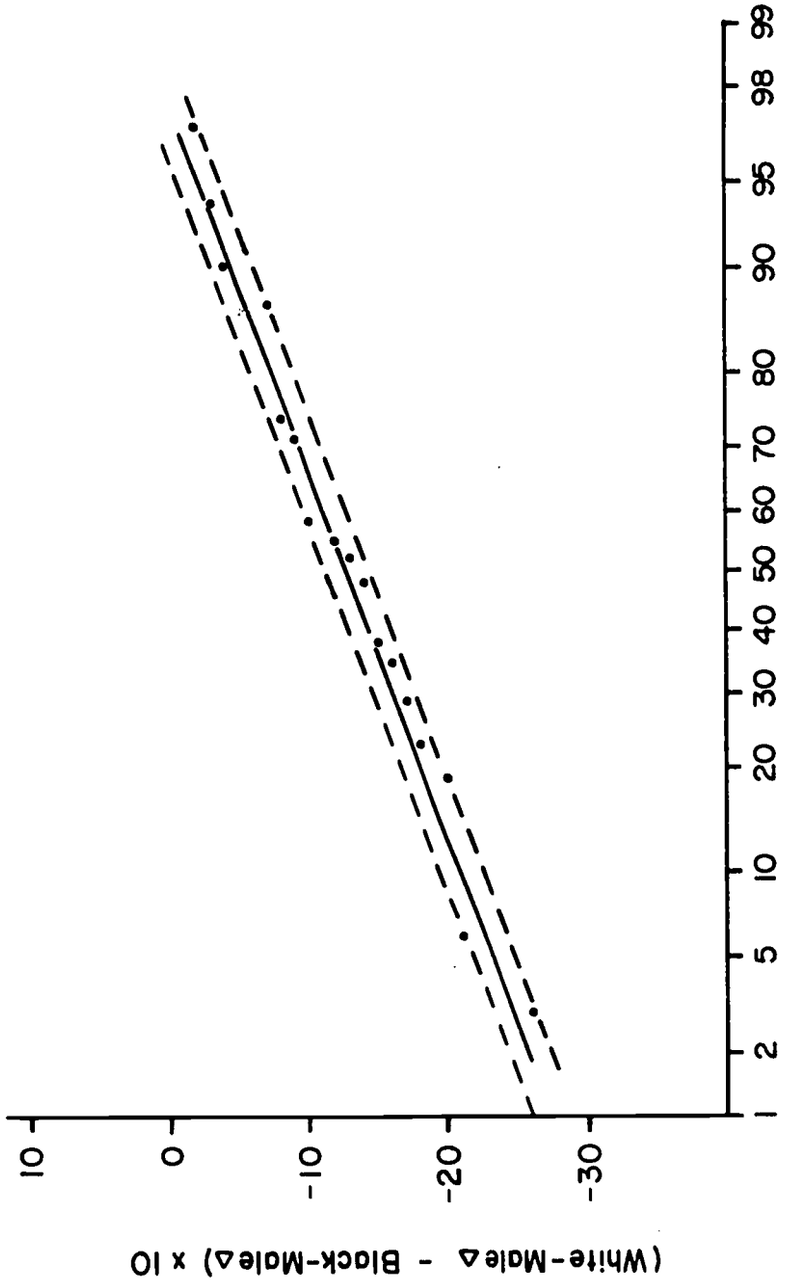

Figure I. White Males vs. Black Males

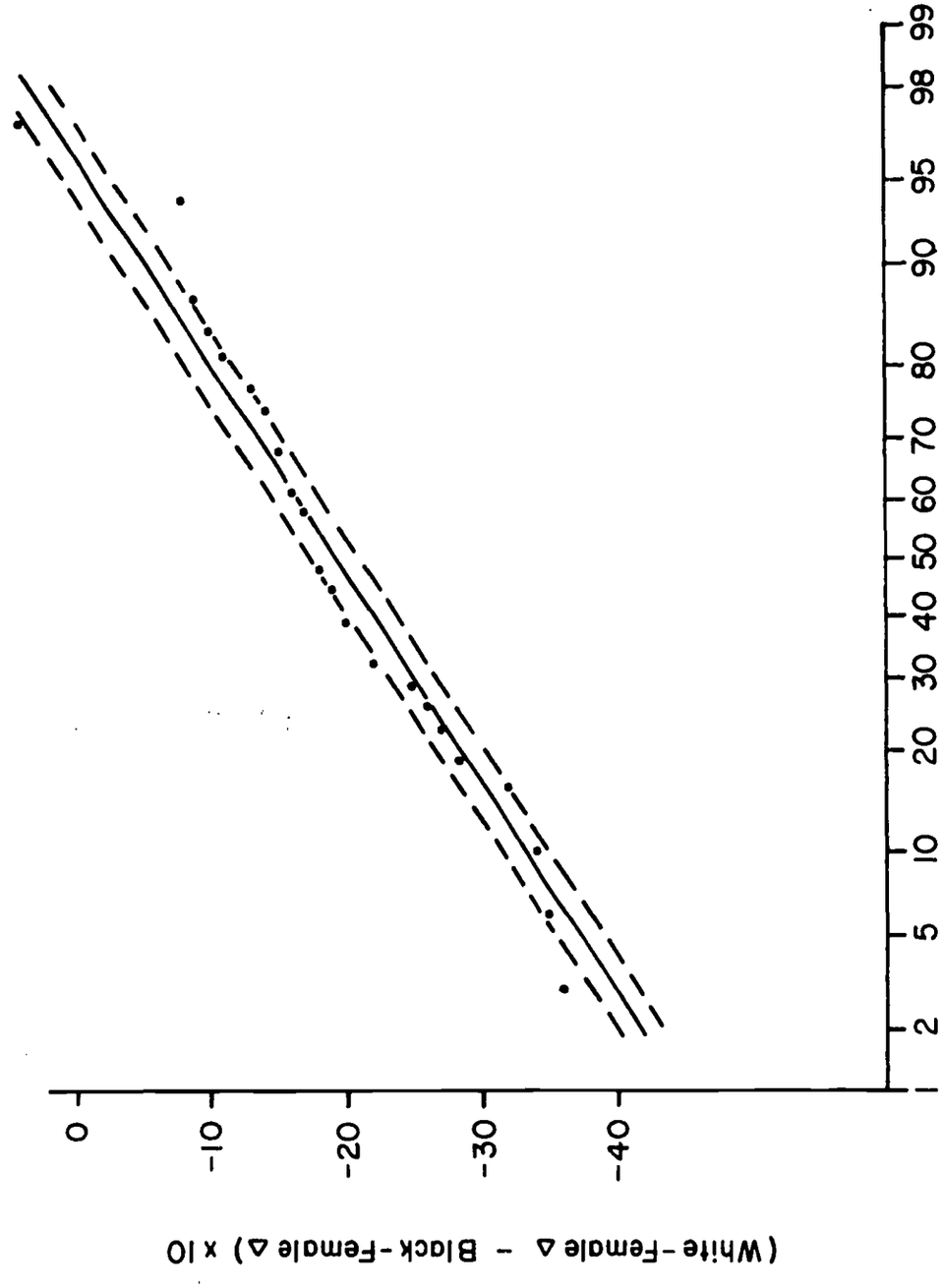

Figure 2. White Females vs. Black Females

Figure 3. Males vs. Females