DOCUMENT RESUME

ED 294 912 TM 011 622

AUTHOR Mellenbergh, Gideon J.
TITLE Experimental Research on the Explanation of Item Bias.
PUB DATE Apr 88
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Arithmetic; Cultural Differences; *Dutch; Elementary School Students; Foreign Countries; Grade 6; Intermediate Grades; Latent Trait Theory; Mathematics Tests; *Minority Groups; *Test Bias; Test Items; *Turkish
IDENTIFIERS *Iterative Methods; *Moroccans

ABSTRACT

Item bias is defined as the dependence of item responses and group membership conditional on the value of the trait that the test is supposed to measure. The results of item bias detection methods based on this conditional definition and using a stepwise or iterative procedure appear to be adequate. In this paper, experimental studies in the Netherlands on the explanation of item bias are reported. For each of the 60 items of an arithmetic test, an assessment was made as to whether the item was biased between Dutch and Turkish/Moroccan students at the end of their sixth-grade year. Hypotheses were formulated to explain the bias. According to the hypotheses, biased items were modified to become less biased and unbiased items were modified to become more biased. The original and modified test versions were randomly administered to each of 169 students of Dutch origin and 93 students of Turkish or Moroccan origin. The statistical tests showed that the hypothesis was confirmed in only three of the 38 cases. (Author/TJH)

Experimental research on the explanation of item bias

Gideon J. Mellenbergh

Department of Psychology, University of Amsterdam, The Netherlands

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 5-9, 1988

Address:
Faculteit der Psychologie,
Universiteit van Amsterdam,
Weesperplein 8,
1018 XA Amsterdam,
The Netherlands

Abstract

Item bias is defined as the dependence of item responses and group membership conditional on the value of the trait that the test is supposed to measure The results of item bias detection methods based on this conditional definition and using a stepwise or iterative procedure appear to be adequate. In this paper experimental studies on the explanation of item bias are reported. For each of the 60 items of an arithmetic test it was investigated whether the item was biased between Dutch and Turkish/Maroccan students at the end of the sixth grade. Hypotheses were formulated to explain the bias. According to the hypotheses biased items were modified to become less biased and unbiased items were modified to become more biased. The original and modified test versions were randomly assigned to each of 169 students of Dutch origin and 93 students of Turkish or Maroccan origin. The statistical tests showed that only in three of the 38 cases the hypothesis was confirmed.

Key Words: Item bias. Iterative Logit Method. definition of item bias. experimental research on the explanation of item bias.

Definition of Item Bias

In item bias research it is investigated whether educational or psychological constructs are differently measured across groups. Item bias research usually starts with the observation that group membership is associated with item responses, e.g the item scores are higher for Whites than for Blacks The situation is shown in Figure 1(a).

Insert Figure 1 about here

The rectangles indicate observed variables. The rectangle denoted Group indicates an observed nominal variable for group membership, such as Black and White, the rectangle denoted Item indicates the observed item responses such as Correct and Incorrect on an Arithmetic item. The double-headed arrow indicates the association between the two variables, e.g. one group tends to have more correct answers on the item than the other group

The finding that Group and Item are associated is, however, not sufficient evidence for the statement that the item is biased. For example, it might be that one group is truly better in arithmetic than the other group, and that, therefore, group membership is associated with the responses on an item measuring arithmetic. This means that a latent Trait, such as latent Artihmetic Ability, is used for explaining the association between Group and Item The situation, where the latent Trait can explain the association between Group and Item, is shown in Figure 1(b). The circle indicates a latent variable and the arrow a causal influence. Trait and Group are associated, i.e one group has lower ability than the other group The item of Figure 1(b) is defined to be unbiased The latent Trait is capable to explain the association between Group and Item. The groups differ in latent ability, but given the level of the latent trait Item and Group are independent. In the literature on contingency table methodology this situation has been called <u>conditional indepence</u> (see, for example, Fienberg, 1980, p. 28): conditonal on the level of the Trait the observed variables Item and Group are independent. In more common language is the Trait the third variable that is responsible for the correlation between the other variables

From the definition of an unbiased item follows immediately the definition of a biased item, given the level of the latent Trait Item and Group are dependent The situation is shown in Figure 1(c).

It is remarked that the definition does not depend on the measurement level of the three variables Usually item bias is described in terms of a dichotomous response variable (e g. correct/incorrect), a nominal group membership variable (e g Black/White), and a latent variable at interval level. But, other types of measurement scales are conceivable, and they fit in the general definition.

For the special case of a dichotomous item response (e.g correct/incorrect), nominal group membership (e.g. Black/White), and a latent trait at the interval level another definition is used: An item is unbiased if its item characteristic curves are identical across groups; otherwise the item is biased. In the special case of a dichotomous response variable, nominal group membership, and an interval latent variable the two definitions are identical (Mellenbergh, 1988)

Item Bias Detection

The main problem in item bias detection is the measurement of the latent trait. Usually a trait is measured using an educational or psychological test In classical psychometrics the total test score is used as an indication of the latent trait, whereas in modern psychometrics the item responses are used for estimating the latent scores. But, in both approaches the same circularity applies. If the test contains biased items the measurement of the latent trait is not free of the bias that is investigated

Several methods for item bias detection have been developed, for a review see Mellenbergh (1988). But, in all methods the above mentioned circularity remains: A biased measure of the latent trait is used for investigating item bias

To break through the circularity Lord (1980, sec 14.5) proposed a stepwise procedure. In the first step the total test score is used for estimating the subjects' latent trait values and for computing item bias statistics. In the second step the biased items are excluded from the test and the reduced test is used for estimating the latent trait values and for investigating item bias Van der Flier, Mellenbergh, Adèr, and Wijn (1984) developed a completely iterative procedure. This so called Iterative Logit Method appeared to be very efficient in detecting simulated biased items (Van der Flier, Mellenbergh, & Adèr, 1984; Van der Flier, Mellenbergh, Adèr, & Wijn, 1984) and in detecting experimentally induced biased items (Kok, Mellenbergh, & Van der Flier, 1985). It is remarked that other item bias detection methods can also be easily extended to iterative procedures and that they might be very efficient as well.

Explanation of Item Bias

In many applications the user is satisfied with the detection of items that are biased with respect to certain groups The items are removed from the test and it is claimed that the test is fair with respect to the groups that have been investigated. But, an important question remains: Why are these items biased? The answer to this question is not only of academic interest, but has also relevance for test construction. If the biasing factors are known the test constructor can prevent the occurence of biased items.

Suppose an item is biased; the bias is graphically represented in Figure 1(c). In this figure, Item and Group are associated as indicated by the double-headed arrow between Item and Group: conditional on the value of the latent trait, Item and Group are dependent, which is the definition of item bias Further, suppose that next to the first trait a second trait is measured by the item. The second trait is an explanation of the item bias when the bias disappears by introducting the second trait. This situation is graphically displayed in Figure 1(d). In Figure 1(c) the item is biased, but in Figure 1(d) the bias has disappeared by introducing the second trait This analysis shows that the search for explanation can be described as "finding the biasing trait(s)" (Mellenbergh & Kok, 1988)

Mellenbergh and Kok (1988) described four research strategies for explaining item bias: (1) qualitative, (2) correlational, (3) quasi-experimental, and (4) experimental In the remainder of this study experiments on the explanation of item bias are reported.

Experiments

The studies were inspired by a similar experiment of Scheuneman (1987) One experiment (Groen, 1988) is completed, whereas the analysis of the data of the second experiment (Molendijk, in preparation) is in progress The first experiment is decribed in some detail.

The test is a 60-item multiple-choice test on arithmetic, administered at the end of primary school in The Netherlands Using the Iterative Logit Method the items were investigated on item bias in a group of 2500 Dutch students and 451 students of Moroccan and Turkish origin at Dutch schools Twelve of the items appeared to be biased between the two groups.

A second version of the test was prepared: The biased items were modified and a number of unbiased items were also modified Some of the biased items were modified more than once and different modifications of the original biased item were included in the second version of the test.

Hypotheses

The items were modified according to one of four hypotheses on the explanation of item bias.

First, it was hypothesized that the plausibility of incorrect options can cause bias For seven unbiased items plausible incorrect options were replaced by less plausible ones; for three biased items less plausible incorrect options were replaced by plausible ones. An example is given in Figure 2.

Insert Figure 2 about here

Second, it was hypothesized that lack of time or fatigue can cause iter bias Four biased items that were at the end of the original test were placed at the beginning of the modified test Four unbiased items at the beginning of the original test were placed at the end of the modified test.

Third, it was hypothesized that the knowledge of words or expressions can cause bias. In six unbiased items words or expressions were replaced by harder words or expressions. In two biased items words or expressions were replaced by easier words or expressions.

Fourth, it was hypothesized that the complexity of the item can cause item bias. For four unbiased items the items were formulated more complex and for eight biased items the items were formulated less complex

Subjects

The test was administered to eighteen schools in Amsterdam. The schools are in neighbourhoods with many Turkish and Maroccan immigrants One of the two versions of the test was randomly assigned to a group of 262 students, consisting of 169 students of Dutch origin and 93 students of Turkish and Maroccan origin.

Data analysis

For each of the items per cell of the 2 (test versions) x 2 (Dutch/Turkish or Maroccan) design the proportion of correct answers was computed An example is given in Table 1. The proportions were analyzed using the logit model (Fienberg, 1980)

Insert Table 1 about here

According to the hypothesis the biased items were modified to become less biased and the unbiased items were modified to become more biased In technical terms this means that in the logit model the interaction of group x test version is of interest. For each of the items the null hypothesis that the interaction parameter is zero was tested at the 5% significance level. Table 1 shows that the difference in the proportions between the two groups is smaller for the modified item than for the original item, which means that the bias has decreased But, the statistical test shows that the effect is not significant at the 5% level.

Results

In total 38 items were modified. In only three of these 38 cases the interaction parameter is significant at the 5% level.

Second experiment

In a second experiment (L. Molendijk) some other hypotheses were tested For example, the hypothesis was tested that the use of decimals in the arithmetic items could cause the bias. The design of this experiment is similar to the design of the first experiment. The only difference is that the same subjects were repeatedly tested: one time the original test was administered and the other time the modified test was applied The data are not yet completely analyzed but the preliminary analyses show the same results as the first experiment: In general the hypotheses are not confirmed.

Discussion

Usually very broad traits are mentioned as explanation of item bias, e.g. the mastery of the item language. In these experiments very specific hypotheses were used: they were formulated at the concrete level of each of the items Our preference is in the direction of rather specific hypotheses because they give more insight in the process that causes the bias.

A disadvantage of a specific hypothesis is, however, that it may be misspecified. Anyway, it appears to be very hard to find the biasing traits.

References

Fienberg, S.E. (1980). The analysis of cross-classified categorical data (2nd ed.) Cambridge, Mass : MIT Press

Groen, H.J. (1988). Vraagpartijdigheid bij rekensommen. [Bias in arithmetic items]. Master thesis, Department of Psychology, University of Amsterdam, The Netherlands.

Kok, F.G., Mellenbergh, G.J., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22, 295-303.

Mellenbergh, G.J. (1988, in press). Item bias and item response theory International Journal of Educational Research

Mellenbergh, G.J., & Kok, F.G. (1988, in press). Finding the biasing trait(s). In S. Irvine, S. Newstead, & P. Dann (Eds.). Computer-based human assessment, Part 3. Identifying group and individual patterns of response to tests. Boston Kluwer-Nijhoff.

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24, 97-118.

Van de Flier, H., Mellenbergh, G.J., & Adèr, H.J. (1984). Een onderzoek naar de effectiviteit van de iteratieve item bias detectie methode bij groepen met een verschillend treknivo. [The effectiveness of an iterative item bias detection method applied to groups differing in mean ability]. Tijdschrift voor Onderwijsresearch, 9, 61-70.

Van der Flier, H., Mellenbergh, G.J., Adèr, H.J., & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

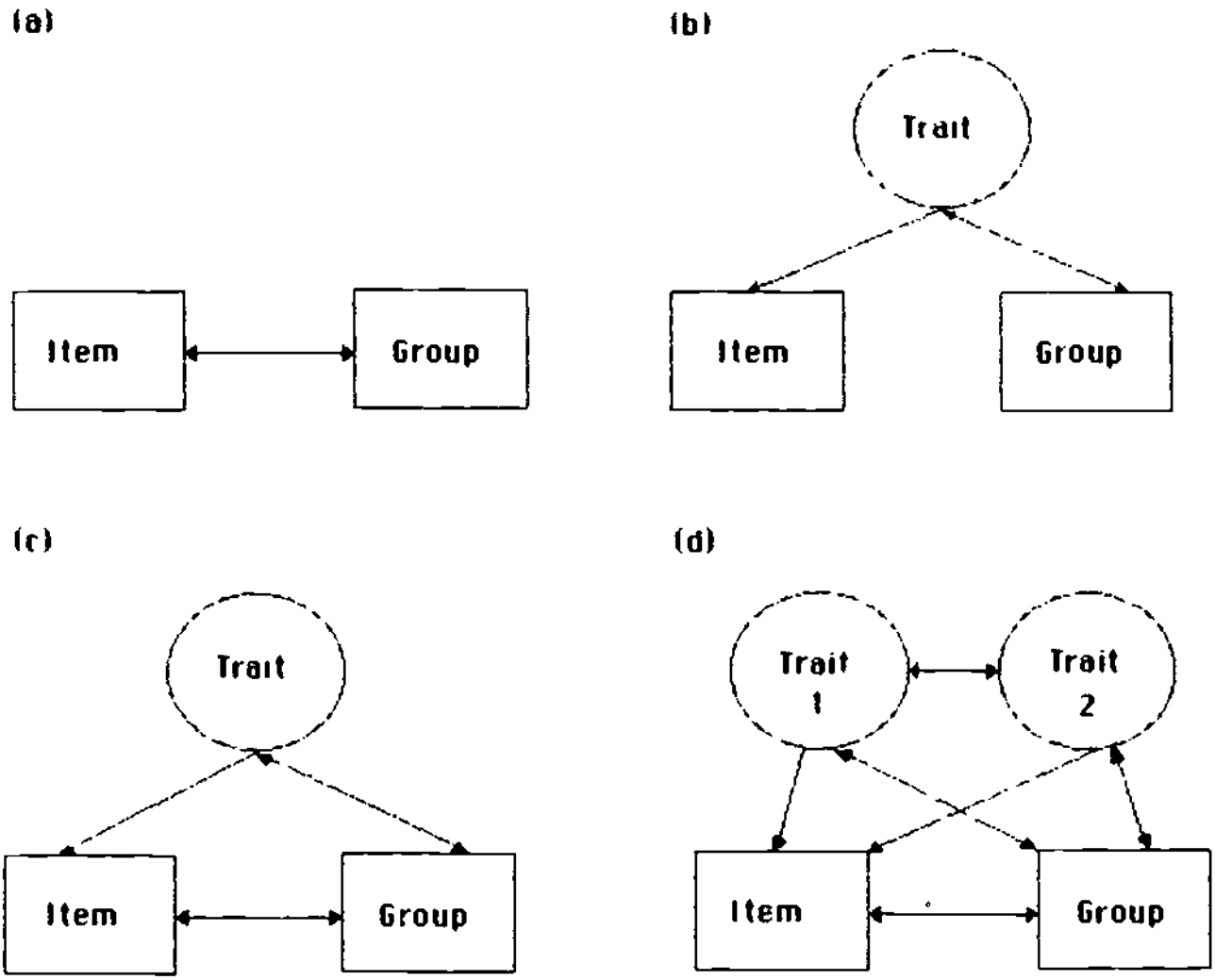

Figure 1 Graphical display of (a) association of group membership and item responses, (b) an unbiased item, (c) a biased item, and (d) a biased item where the bias disappears by introducing an additional trait.

Original Item (Biased)	Modified Item
457 - 2,34 =	457 - 2,43 =
A. 2,23	A. 453,69
B. 454,66	B. 454,66
C. 454,76	C. 454,76
D. The correct answer is not given	D. The correct answer is not given

Figure 2 Example of a biased item (no. 33) where the less plausible option A is replaced by a more plausible one.

Table 1

Proportion of correct answers per cell of the 2 (versions) x 2 (groups) design, item no. 33

Test version	Group	
	Dutch (N=169)	Turkish/Moroccan (N=93)
Original (Biased)	.60	.40
Modified	.63	.53