DOCUMENT RESUME

ED 194 533 TM 800 659

AUTHOR	Linn, Robert L.; Harnisch, Delwyn L.
TITLE	Interactions Between Item Content And Group Membership on Achievement Test Items.
PUB DATE	Apr 79
NOTE	22p.: Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979).
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Achievement Tests; *Black Students; Grade 8; Item Analysis; Junior High Schools; *Latent Trait Theory; *Mathematics; Scores; *Test Bias; *Test Results
IDENTIFIERS	Illinois Inventory of Educational Progress; Rasch Model

ABSTRACT

The purpose of this investigation was to examine the interaction of item content and group membership on achievement test items. Estimates of the parameters of the three parameter logistic model were obtained on the 46 item math test for the sample of eighth grade students (N = 2055) participating in the Illinois Inventory of Educational Progress, Illinois' statewide assessment booklet. Black students were divided into quintiles on the basis of their estimated theta's. Average standardized difference scores were computed within each quintile as well as across all black students. Some items were identified as biased in favor of black students because their actual performance is systematically better than expected while some were considered biased against black students relative to other items on the math test. Finally, items were classified by content categories and compared in terms of their average standardized difference score. Questions about the metric system, those involving definitions, and those based upon graphs and figures stood out as ones on which black students did worse than expected from the latent trait model. Story problems involving money, those involving unknown symbol substitution stood out in the other direction. Recommendations are made for test revision and for a line of inquiry into what is now labeled as item bias. (Author)

Interactions Between Item Content And Group Membership on Achievement Test Items

Robert L. Linn and Delwyn L. Harnisch
University of Illinois, Urbana-Champaign

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1979

Abstract

The purpose of this investigation was to examine the interaction of item content and group membership on achievement test items. Estimates of the parameters of the three parameter logistic model were obtained on the 46 item math test for the sample of eighth grade students (N = 2055) participating in the Illinois Inventory of Educational Progress, Illinois' statewide assessment booklet. Black students were divided into quintiles on the basis of their estimated theta's. Average standardized difference scores were computed within each quintile as well as across all black students. Some items were identified as biased in favor of black students because their actual performance is systematically better than expected while some were considered biased against black students relative to other items on the math test. Finally, items were classifeid by content categories and compared in terms of their average standardized difference score. Questions about the metric system, those involving definitions, and those based upon graphs and figures stood out as ones on which black students did worse than expected from the latent trait model. Story problems involving money, those involving unknown symbol substitution stood out in the other direction. Recommendations are made for test revision and for a line of inquiry into what is now labeled as item bias.

Interactions Between Item Content and Group Membership on Achievement Test Items

It has been documented repeatedly that "...children of the poor tend to score lower...on standardized ability and achievement tests than do children of the affluent..." (Stanley, 1971, p. 640). Similarly, and not unrelated, many studies have shown that members of some minority groups tend to score lower on standardized ability and achievement tests than do members of the white majority. There are a number of other variables that could also be used to identify subgroups of children that tend to have relatively poor performance on standardized tests. Some of these have already been discussed in this symposium. In particular, groups of children who differ in terms of motivation, test-taking anxiety, or in their success and failure attributions also differ in their average performance on achievement tests. Performance differences are also to be expected for groups that have had differential exposure to the content.

By themselves, of course, group differences in performance on a test tell us nothing about the reasons for those differences. The differences alone cannot even distinguish between the possibility that the results are reflecting real and important differences in the knoweldge and understanding of the subject matter covered by the test and the possibility that the results reflect inadequacies in the measuring instrument. In other words, the test might be biased against one of the groups. Anastasi has cogently argued that: "No test can eliminate causality. Nor can a test score, however derived, reveal the origin of the behavior it reflects" (1961, p. 389). Nonetheless, for test results to have any utility they must be interpreted.

When scores on a mathematics achievement test are interpreted as indicators of competence in mathematics, a host of alternative explanations must be ruled out either explicitly, or as is more often the case, implicitly. The evaluation of alternative explanations of test performance often involves the investigation of possible interactions. For example, an interaction between speediness of a test and student anxiety (Hill, 1977) may indicate that a test given under its standardized timing conditions yields biased estimates of the competence of high anxious children. Ken Hill has already discussed some interactions between student characteristics and test taking conditions and we will not dwell further on this type of interaction. Our focus is on another type of interaction; one between characteristics of items and characteristics of examinees.

It is often argued that certain types of items are biased against some groups. Of particular concern is the possibility that non-essential characteristics of particular test items may result in misleadingly poor performance for minority and/or socioe'onomically disadvantaged children. For example, when vocabulary is incidental tc the skill that the items are purported to measure (e g. an arithmetic story problem), then the use of words that are less familiar to members of one group than to another may result in a biased indication of the relative performance of the two groups.

There are a variety of student characteristics that might interact with item characteristics in ways that affect overall performance on a test. Ethnic group membership or socioeconomic status are but two of many potentially important characteristics. Differences in motivation or in test taking anxiety could also lead to interactions with characteristics of test items. The identification and understanding of possible interactions between student characteristics and the characteristics of items used to measure student achievement could contribute to the development of improved measurement procedures.

Ideally, the possibility of characteristics of items that interact with student characteristics would be investigated experimentally. Item characteristics such as socioeconomic status or level of anxiety would be systematically varied and compared experimentally. A study by Medley and Quirk (1974) is an example of this approach. Medley and Quirk used altered content specifications for the general education items of the National Teacher Examinations (NTE). For two experimental forms of the examination, the proportion of items involving contemporary culture (modern items) and the proportion of items involving black cultural contributions (black items) were increased and the proportion of items dealing with classical contributions (traditional items) was reduced. The relative performance of black and white candidates was then compared on the three types of items. The black candidates did relatively better on the black and modern items than on traditional items, and when black and modern items were compared their relative performance was better on the black items.

For many types of achievement tests, the item characteristics that might interact with student characteristics are less obvious. For example, characteristics of arithmetic items that might interact with student characteristics are less apparent than those of items dealing with cultural contributions such as Medley and Quirk investigated. Consequently, a more exploratory approach is needed to try to identify unsuspected item characteristics that may tend to yield misleading results for particular subgroups of examinees.

The idea of identifying item characteristics that eventuate in the underestimation of the competence of identifiable groups of students is not a new one. It was the dominant idea of the landmark study of Eells, Davis, Havighurst, Herrick, and Tyler (1951). Their stated purpose was to "identify (a) those kinds of test problems on which children from high socioeconomic backgrounds show the greatest superiority and (b) those kinds of test problems on which children from low sociocconomic backgrounds do relatively well" (p. 6). In the late 1960's and early 1970's, a number of studies (e.g. Angoff & Ford, 1973; Cleary & Hilton, 1968; Green, 1973; Jensen, 1974) were conducted to identify items that are unusually difficult or that function differently for members of a particular minority group.

The previous efforts have not been overly successful in identifying general characteristics of test items that result in the underestimation of the competence of particular groups of students. A possible reason that the results were not more informative is that the early studies relied on sample dependent item statistics. Furthermore, the most commonly used item statistic, item difficulty, is confounded with other item charactistics such as item discriminating power (see Hunter, 1975). As suggested by Lord (1977) and Wright (1977), latent trait theory provides a theoretically sounder approach to the problem of identifying items that function differently for different groups. A few recent studies (e.g. Harms, 1978; Ironson, 1978; Rudner, 1977) have compared the use of latent trait models to identify biased items to a variety of techniques that had previously been used. While some agreement among the techniques was found, it is clear that the more commonly used techniques are not substitutes for an approach based on latent trait theory.

One of the great appeals of latent trait models is the possibility that they can provide invariant item parameters. That is, in the words of Wright, "person-free item calibration." If a given latent trait model holds, then the item parameter estimates obtained separately for two different groups should be the same, except for sampling error, once they have been put on the same metric.

With a sufficiently large sample size in each group, our preference would be to use the three-parameter logistic model (Birnbaum, 1968). Estimates would be obtained separately for each group and placed on the same scale by a linear equating of the difficulty parameter estimates. Comparisons of the item characteristic curves would be made along the lines outlined by Lord (1977), Ironson (1978) and Warm (1987). Unfortunately, this approach requires very large samples. Following Lord's guidelines for the use of LOGIST, a sample of 1000 would be needed in the smallest of the groups being compared.

In our research, we are interested in making comparisons of a variety of types of subgroups. The size of the smallest group is often around two or three hundred and sometimes as small as one hundred. Consequently, an alternative approach was needed.

One alternative is, of course, to use a simpler model, namely the one parameter Rasch model. With this model items would be separately calibrated for each group. The estimates of item difficulty would be placed on a common metric by means of a linear transformation that equates the mean and standard deviation of Rasch difficulties for one group to those of another. Differences in difficu.ty of an item for two groups would then be used to determine the direction and degree of "bias".

A concern in the use of the Rasch model is that group differences in difficulty estimates may be an artifact of item differences in discriminating power or location of the lower asymptote. Thus, we decided to use a second approach. We first obtained estimates of the item discriminating power, a, the item difficulty, b, and the lower asymptote, c, of the three-parameter logistic model based upon all available cases in the sample. This also provided an estimate of each person's location along the latent trait, θ. From these estimates, Pij, the estimated probablity that person j would answer item i correctly, were obtained in the usual way. That is,

$$P_{ij} = c_i + \frac{1 - c_i}{1 + \exp\,[-1.7a_i\ (\theta_j - b_i)]} ,$$

where a_i, b_i, c_i and θ, are all estimates. These estimates based on the model when averaged over members of a subgroup can be compared to the observed proportion correct for that subgroup. If person j is a member of group g, then the proportion of people in group g expected to get item i correct according to the model is

$$P_{i.} = \frac{1}{n_g} \sum_{j\varepsilon g} P_{ij}$$

when n_g is the number of persons in group g. The observed proportion correct on item i for group g, O_i, is simply the number of people in group g who answer item i correctly divided by n_g. The difference,

$$D_i = O_i - P_{i}. ;$$

is an index of the degree to which members of that subgroup perform better or worse than expected on that item. We have used D_i and a standardized difference between observed and expected performance for members of various subgroups to identify items that are unusually easy or difficult for members of those subgroups.* Items in these categories are finally compared in terms of item content and format.

Results for Black Students in Eighth Grade Mathematics Test

In the time available, it is impossible to report on the results of all tests at all three grade levels included in the Illinois Inventory of Educational Progress. Nor is it possible to give results for the variety of subgroups that we have investigated and are continuing to investigate. Instead we have decided to focus on one test, mathematics, at one grade level, eighth, for one subgroup, black students. Thus, the results may be taken as illustrative of the type that may be obtained on other tests, at other grade levels and for other groups.

The eighth grade mathematics test contains 46 items. Several types of items are included. There are straight forward calculation problems, e.g. 1/2 + 1/3 = ?, there are story problems, problems involving substitution for an unknown or the solution of an equation for an unknown. There are also questions involving definitions, graphs, and the metric system. The apparent differences are enough that the undimensionality of the items that is assumed by the latent trait models is only crudely approximated. On the other hand, the variability is in line with what is found on many achievement tests

*The standardized difference score is

$$Z_j = \frac{1}{n_g} \sum_{j \varepsilon g} \left[\frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}\,(1 - P_{ij})}} \right]$$

where U_{ij} = 1 person j answersitem i correctly and U_{ij} = 0 otherwise.

that result in a single total-mathematics score. Hence, it is of potential importance to see if items that differ in their content and format characteristics show systematic differences in the direction and magnitude of the observed vs. the expected proportion correct (i.e. either D_i or Z_i) for particular subgroups.

The estimates of the parameters of the three parameter logistic model were obtained for the sample of all eighth grade students with usable data (N = 2,055). One item was deleted from the analysis because the estimate of discriminating power approached zero. All results are based on the remaining 45 items. These results were then used to obtain estimates of the expected chance that each student would have of getting each item right, P_{ij}. These estimates were compared to the observed results separately for white and for black students. The comparisons for the 283 black students in the sample were made as a function of estimated values of θ. First the black students were divided into quintiles on the basis of their estimated θ's. Within each quintile, $P_{i.}$, O_i and D_i was then computed for each item. This allowed for the possibility that students at one level of θ may perform systematically better than expected on an item while those at another level perform systematically worse than expected.

Average standardized difference scores, Z_i, were also computed within each quintile as well as across all black students. The latter was used as an overall index for an item.

The expected and observed proportion correct for the five quintiles of four items are shown in Figures 1 and 2. The points, P_i and O_i are plotted above the mean theta value within each quintile. The solid lines connect the expected proportions, $P_i.$, and the dashed lines connect the observed proportions. The two items shown in Figure 1 have the largest positive standardized difference score averaged over all quintiles. The items depicted in Figure 2 have the largest negative standardized difference score averaged over all quintiles.

Relative to other items in the tests, the items in Figure 1 may be considered biased in favor of black students because their acutal performance is systematically better than expected. The items in Figure 2, on the other hand, may be considered biased against black students relative to other items on the test.

The results in Figures 1 and 2 represent the extremes. For these items, however, the differences are fairly substantial. The performance of black students would appear better on a test that had more items of the type shown in Figure 1 and/or fewer items of the type shown in Figure 2. But, these results do not, of themselves, indicate what leads to these differences. We have attempted to find clues as to the possible reasons for these differences in two ways. First, we simply compared the content and format of the items with the most extreme positive differences with those with the most extreme negative differences. Second, we categorized items in several ways based on their content and format and then compared the items in different categories in terms of their average standardized difference scores, Z_i.

Of the five items with the largest negative standardized differences, three of them involved questions about the metric system while the other two involved definitions (e.g. "An angle may be measured in units called: 1. centimeters, 2. degrees, 3. grams, 4. inches"). (Both items depicted in Figure 2 involve questions about the metric system.) In contrast, none of five items on which the actual performance of black students was better than expected by the greatest amount (as measured by the largest Z_i's) involved the metric system or definitions. Two were calculation items, one was a substitution for an unknown and calculation ($X^2 - 1 = ?$ where $X = 3$), and the other two were story problems. (One of the items depicted in Figure 1 is a calculation problem and the other a story problem.) These results are suggestive, but hardly conclusive.

When items were categorized by content, the patterns of standardized difference scores were quite distinctive for some of the categories. Of the six items involving the metric system, five of them had negative Z_i's and the sixth one was zero. Similarly, 7 of the 8 definition questions had negative Z's, while the remaining one had a small positive value. At first inspection the results for story problems were mixed. When these items were divided into questions dealing with money vs. others, however, the results appeared more consistent*. The black students did better than expected on all 5 items dealing with money. They did worse than expected from the latent trait model results, however, on 5 of the 8 remaining story problems. With two exceptions, the differences on the 12 calculation problems were small (Z_i between -.06 and +.07). Black students performed better than predicted from the model on the two exceptions (Z_i = .15 and .23).

The mean Z_i and range of Z's within each category of items are listed in Table 1. Questions about the metric system, those involving definitions and those based upon graphs and figures stand out as ones on which black students

*An example of a story problem dealing with money is:

> "Television sets are on sale at two stores. One offers a ten percent discount while the other offers 15 percent. What is the difference in the sale price at the two stores of a TV set that is regularly priced at $100?"
>
> 1. $5
> 2. $10
> 3. $15
> 4. $20

did worse than expected from the latent trait model. Story problems involving money, those involving unknown symbol substitution and, to a lesser extent, calculation problems stand out in the other direction. That is, black students tend to do better than expected from the model on those types of problems.

Discussion

Differences such as those reported above might be labeled item bias. But, to say that questions about the metric system are "biased" against black students and story problems involving money are "biased" in their favor is not very helpful. It implies that the items are at fault. It is at least as plausible however, that the model is at fault and/or that the "bias" is due to instructional differences. The assumption of unidemsionality is clearly violated for this set of items. Hunter (1975) has shown that items may appear "biased" using latent trait models as a result of violations of the unidimensionality assumption.

The observation that the results may be due to multidimensionality does not detract from their potential utility. It seems clear that global scores on most survey tests of achievement are based on items that reflect more than one dimension, albeit not necessarily to the degree that this is true of the test we analyzed. As long as this is true and subsets of items show consistent differences of the type we have illustrated, then the magnitude of group differences on the global score will depend on the number of items in various categ Deleting or reducing the number of questions about the metric system or increasing the proportion of story problems that involve monetary calculations would be expected to alter the magnitude of group differences on the global score in predictable ways.

The identification of types of items that function differently for different groups is only a first step. It leaves unanswered the more interesting question of why. We have some speculations, but additional work is needed to provide any support for them. The seemingly most natural speculation is that the amount of instruction in areas reflected by the various categories is not the same for black as for white students. Differences in instructional patterns could result from attendance patterns and school to school variability in content coverage and emphasis. We are currently exploring this possibility. Our goal is to classify schools according to teacher questionnaire responses regarding student exposure to material judged necessary to answer particular items. The comparison of observed and expected performance of students attending a homogeneous category of schools would then be compared along the lines used for the results that we have just presented. With such an analysis we expect that what may appear now as "item bias" might better be labeled "instructional bias."

References

Anastasi, A. Psycholgocal tests: Uses and abuses. *Teachers College Record*, 1961, *62*, 389-393.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, *10*, 95-106.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968, Chapters 17-20.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, *28*, 61-75.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. *Intelligence and cultural differences*. Chicago, Ill: Chicago Press, 1951.

Green, D. R. *Racial and ethnic bias in achievement tests and what to do about it*. Monterey, California: CTB/McGraw-Hill, 1973.

Harms, R. A. *A comparative concurrent validation of selected estimators of test item bias*. Unpublished doctoral dissertation, University of South Florida, 1978.

Hill, K. T. The relation of evaluative practices to test anxiety and achievement related motives. *Educator*, 1977, *19*, 15-21.

Hunter, J. E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, 'laryland, December 2-5, 1975.

Ironson, G. H. A comparative analysis of several methods of assessing item bias. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, 1978.

Jensen, A. R. How biased are culture-loaded tests? *Genetic Psychology Monographs*, 1974, *40*, 185-244.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Medley, D. M., & Quirk, T. J. The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examination. Journal of Educational Measurement, 1974, 11, 235-245.

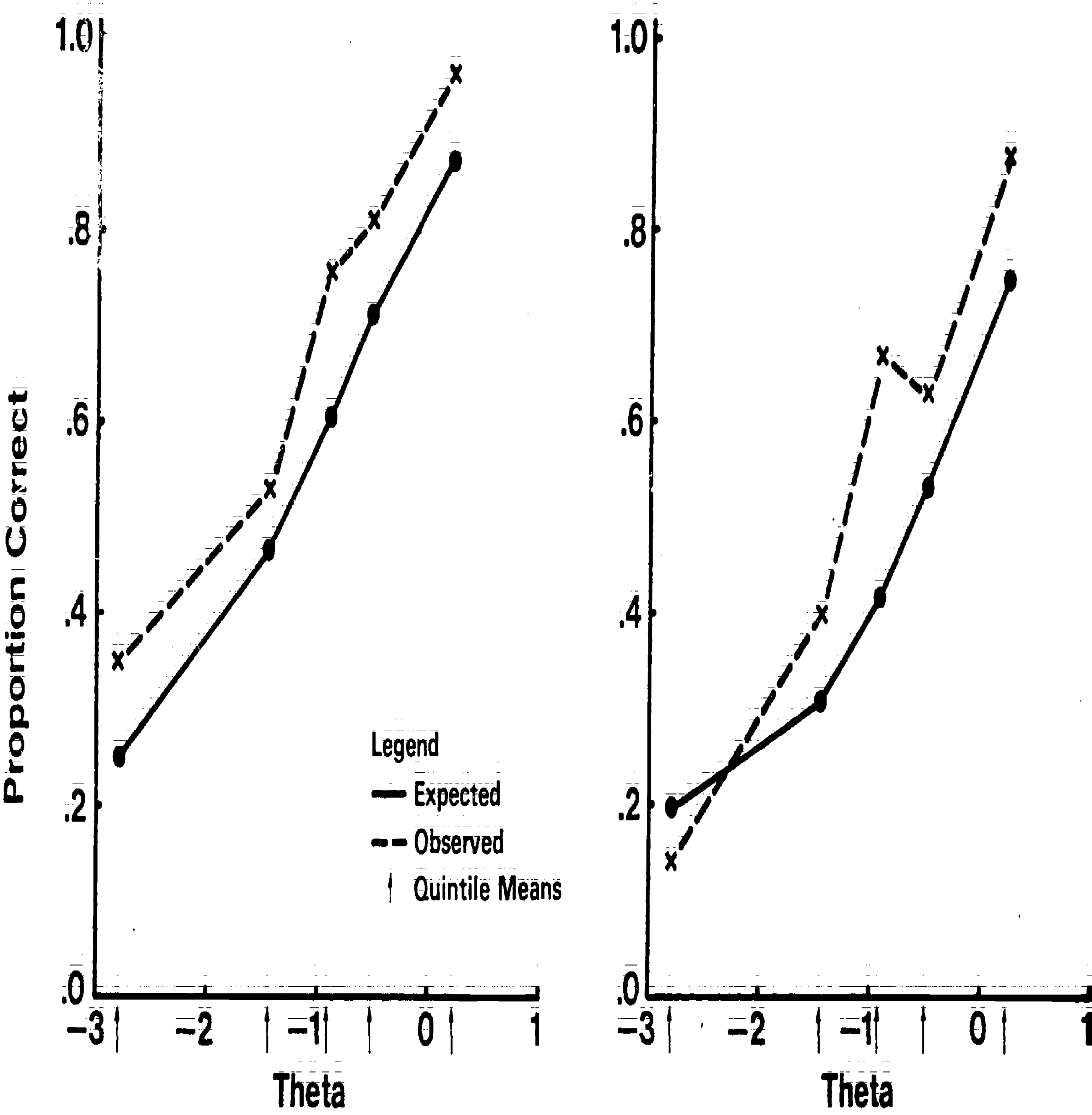
Rudner, L. M. An evaluation of select approaches for biased item identification. Unpublished docotral dissertation, Catholic University of America, 1977.

Stanley, J. C. Predicting college success of the educationally disadvantaged. Science, 1971, 171, 640-647.

Warm, T. A. A primer of item response theory. Department of Transportation Technical Reort 941078, U.S. Coast Guard Institute, Oklahoma City, 1978.

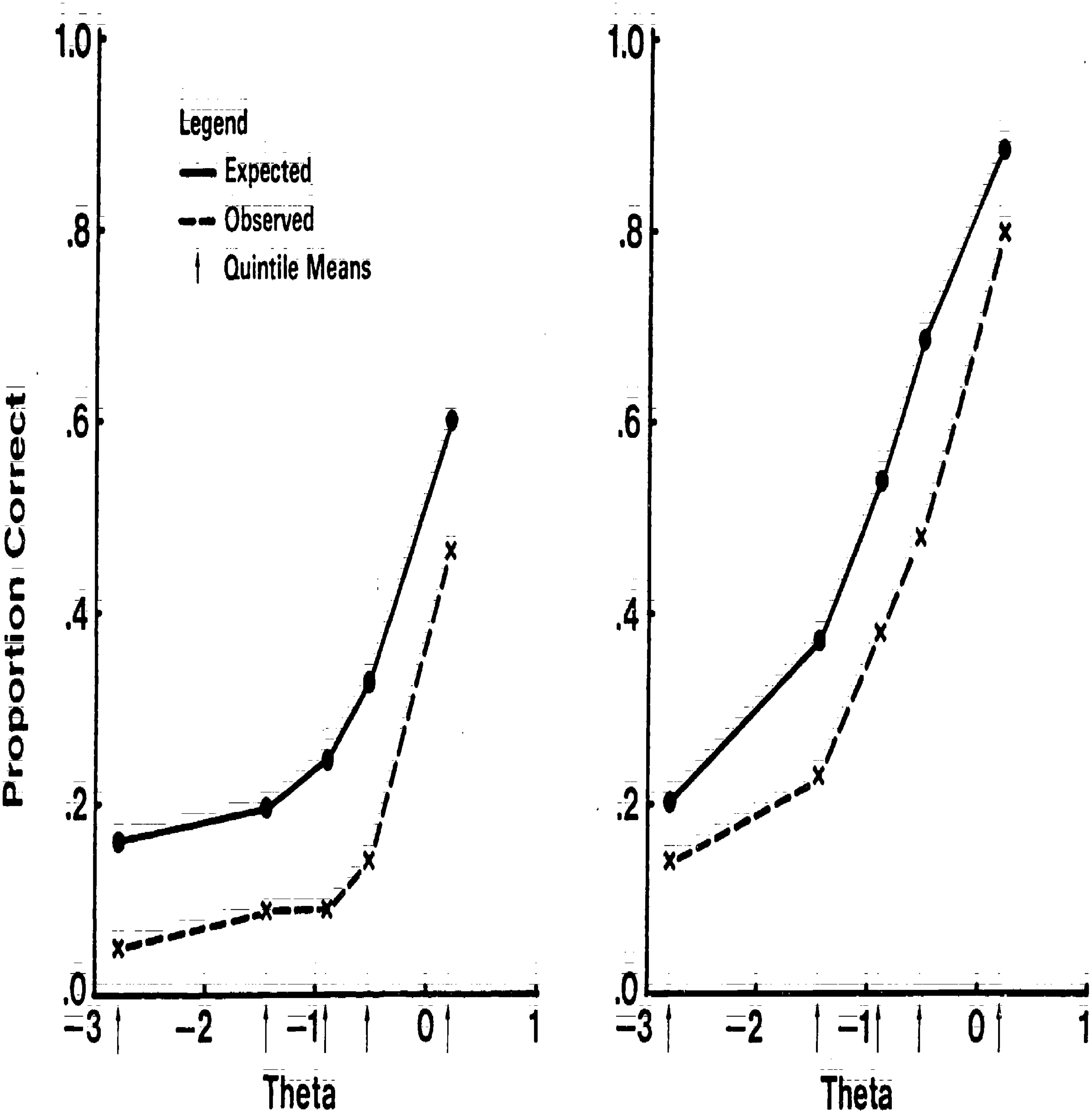
Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

FIGURE 1



Observed and Expected Proportion Correct as Function of Theta for Two Items with Large Positive Differences

FIGURE 2



Observed and Expected Proportion Correct as Function of Theta for Two Items with Large Negative Differences

Table 1

The Mean and Range of Standardized Differences
between Observed and Expected
Proportion Correct by Content Categories

Category	Number of Items	Mean Z	Range of Z's
Calculation	12	.04	-.06 to .23
Definitions	8	-.08	-.20 to .07
Story Problems (General)	8	-.05	-.14 to .09
Story Problems (Money)	5	.13	.05 to .21
Metric System	6	-.18	-.32 to .00
Graphs and Figures	3	-.08	-.11 to -.06
Unknown Symbol Substitution	2	.08	.03 to .14
Unclassified	1	.12	.12