DOCUMENT RESUME

ED 395 001 TM 024 997

AUTHOR	Allen, Nancy L.; Wainer, Howard
TITLE	Nonresponse in Declared Ethnicity and the Identification of Differentially Functioning Items. Program Statistics Research, Technical Report No. 89-89.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-89-47
PUB DATE	Oct 89
NOTE	22p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Blacks; Comparative Analysis; *Ethnicity; Identification; *Item Bias; *Minority Groups; Racial Identification; *Responses; *Test Items; Test Results; Validity; Verbal Tests; Whites
IDENTIFIERS	Item Bias Detection; *Mantel Haenszel Procedure; *Nonresponders; Scholastic Aptitude Test; Self Report Measures

ABSTRACT

The accuracy of procedures that are used to compare the performance of different groups of examinees on test items obviously depends on the correct classification of members in each examinee group. The significance of this dependence is determined by the sensitivity of the statistical procedure and the proportion of examinees who are unidentified. Since the number of nonrespondents to questions of ethnicity is often of the same order of magnitude as the number of identified members of most minority groups, understanding the effect of nonresponse is crucial to evaluating the validity of a procedure that is used to study differential item functioning (DIF). This study examined the effect of nonresponse to questions of ethnic identity on the measurement of DIF for Scholastic Aptitude Test verbal items using a commonly used modern method, the Mantel-Haenszel procedure (P. W. Holland and D. T. Thayer, 1988). The study considered 94,486 examinees who identified themselves as black or white or did not identify themselves by ethnic group. It was found that efforts to obtain more complete ethnic identifications from the examinees would be rewarded with more accurate DIF analyses. (Contains four tables and eight references.) (Author/SLD)

RR-89-47

Nonresponse in Declared Ethnicity and the Identification of Differentially Functioning Items

Nancy L. Allen

and

Howard Wainer

Educational Testing Service

PROGRAM
STATISTICS
RESEARCH

TECHNICAL REPORT NO. 89-89

BEST COPY AVAILABLE

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

NONRESPONSE IN DECLARED ETHNICITY AND THE IDENTIFICATION OF DIFFERENTIALLY FUNCTIONING ITEMS

Nancy L. Allen

and

Howard Wainer

Educational Testing Service[1]

Program Statistics Research
Technical Report No. 89-89

Research Report No. 89-47

Educational Testing Service
Princeton, New Jersey 08541-0001

October 1989

[1] This research was conducted while the first author was an ETS postdoctoral fellow. It was funded by the Educational Testing Service. We are gratef"l to Bill Angoff, Henry Braun, Vic Bunderson and Neil Dorans for their insightful comments and suggestions. Nancy Burton, Bob Cameron, Nancy Robertson and Nancy Wright provided us with information.

TABLE OF CONTENTS

Page

1. INTRODUCTION.............................. 1

2. DATA AND METHODS.......................... 4

3. DISCUSSION AND CONCLUSIONS................ 10

REFERENCES.................................... 12

TABLE 1: Ethnic Identification................ 13

TABLE 2: Ethnic classifications for each analysis............................ 14

TABLE 3: Item identification by the ETS Mantel-Haenszel Guidelines........... 15

TABLE 4: Sample Sizes........................ 16

Abstract

The accuracy of procedures that are used to compare the performance of different groups of examinees on test items obviously depends upon the correct classification of members in each examinee group. The significance of this dependence is determined by the sensitivity of the statistical procedure and the proportion of examinees who are unidentified. Since the number of nonrespondents to questions of ethnicity is often of the same order of magnitude as the number of identified members of most minority groups, understanding the effect of nonresponse is crucial to evaluating the validity of a procedure which is used to study differential item functioning (DIF).

In this study, we examined the effect of nonresponse to questions of ethnic identity on the measurement of DIF for SAT verbal items using a commonly used modern method, the Mantel-Haenszel procedure (Holland and Thayer, 1988). We found that efforts to obtain more complete ethnic identifications from the examinees would be rewarded with more accurate DIF analyses.

I. INTRODUCTION

In the past, little emphasis has been placed on the variability of measures used to compare the performance of items for different groups of examinees, except in terms of sampling variance (e.g. Phillips and Holland, 1987). Recently, however, there has been interest in variability due to threats to the generalizability of these measures. For instance, Holland and Longford (N. Longford, personal communication, March 10, 1989) are presently studying the variation of an item bias measure for the same test administered to two populations of examinees (as defined by different assessment years). A source of variation in item bias measures that may be a threat to generalizability, but which has not been examined previously, is the lack of information about subgroup membership for some examinees. A lack of information about subgroup membership may mean that the value of an item bias measure can not be generalized to the situation when all examinees are identified properly.

For instance, examinees who take the SAT have the option of filling out a separate form, the Student Descriptive Questionnaire (SDQ), on which they are asked to identify their ethnicity. Despite the fact that a few questions (such as one requesting gender identification) appear on the answer form, the SDQ is the only way that ethnic information is obtained. The percentage of examinees who chose not to complete the SDQ in 1985, for example, was about 14% (about 138,000 examinees). With the introduction of a new form of the SDQ in 1987, this percentage has been lowered considerably (to about 5%). However, the missing information makes the "nonresponders to the ethnicity question" the second largest ethnic designation (behind "White"). If inferences are to be made about the performance of various ethnic groups who take the SAT, they must be made only on the basis of those who choose to identify themselves. To make inferences that generalize to an entire et'hnic group we must assume that for each group, examinees who identify their ethnicity are no different, in terms of their performance, than those who do not identify their ethnicity. This assumption is usually called ignorable nonresponse (Little and Rubin, 1987). If nonresponse is not ignorable, any inferences we make concerning ethnic performance would be inaccurate. It would not be solely test or item performance that we would be describing, because performance would be confounded with the inclination to fill out the SDQ. Therefore, an important question is, "To what extent are measures of ethnic bias of items affected by nonresponse to the ethnicity question on the SDQ?"

In order to be fully aware of the problems int. 'uced by nonresponse to the SDQ for accurate estimates of item performance witnin separate subpopulations of examinees, it is important to understand how methods that examine item bias depend on an examinee's ethnicity. While the examination of items for ethnic bias is now a part of standard test construction practice and recent research proliferates on the topic, the methods for doing this are not standardized (see Shepard, Camilli and Williams, 1985, for a review of some traditional approaches; and Dorans and Kulick, 1986, and Thissen, Steinberg and Wainer, 1988, for more recent developments). Because of its recent impact on test development, we will examine the Mantel-Haenszel (M-H) procedure, a procedure that was developed in the late 1950's by two biostatisticians (Mantel and Haenszel, 1959) as a method for evaluating the efficacy of cancer treatment. It was adapted to the study of item functioning by Holland and Thayer (1988). This method looks for differences in performance only within comparable groups and reveals what is commonly called "differential item functioning," or DIF, a term that is more precise and less value-laden than "item bias."

The Mantel-Haenszel procedure begins by dividing the examinee population into a reference group whose performance is considered to be the standard (often, the majority in the population), and a focal group, whose performance is to be compared with the reference group's (typically, the minority group). The two groups are stratified by an overall performance variable related to their proficiency (usually the total score on the test) and then a M-H statistic based on the common odds ratio of the number of examinees who successfully answered the question in the reference group compared to the number in the focal group (M-H D-DIF; Holland and Thayer, 1988) is calculated. A statistically significant M-H D-DIF statistic implies the existence of DIF. It is common practice that items that the two groups find differentially difficult are set aside for special treatment.

At this time the Educational Testing Service (ETS) categorizes items to be used in most tests in three ways.[2] "Type a" items are items that are used first in test development, because there is no evidence to suggest that these items are performing differently for different groups after proficiency is taken into account. "Type b" items are used next, if necessary, to fulfill the test specifications. These are items for which the M-H D-DIF statistic is in the mid-range of values. "Type c" items have extreme D-DIF values and, so, all type c items found in pretesting are subjected to careful scrutiny. If they are found to fail content oriented criteria they are deleted from the item pool. Type c items that are found in operational testing which cannot be justified on content

[2] Current ETS practice in categorizing an item is somewhat complex. If the absolute value of M-H D-DIF for the item is less than 1 or is not significantly different from 0 (at the .05 level), then the item is classified as a "type a" item. If the absolute value of D-DIF is between 1 and 1.5 or is at least 1 but not significantly greater than 1 (at the .05 level), it is classified as a "type b" item. If it is at least 1.5 and significantly greater than 1 (at the .05 level) then the item is classified as a "type c" item.

grounds are deleted from the test and not scored. Two notes should be made about using the Mantel-Haenszel guidelines in this study. First, the criterion for the M-H procedure is based on the number right score, while the test is actually scored using formula scores. Items identified in this study may contribute differently to formula scores than they do to number right scores. Second, the sample of examinees in this study is different from the samples used in practice to make decisions about items. The examinees in this study have all taken the SAT on two very specific occasions.

In terms of the Mantel-Haenszel procedure, nonresponse to the ethnicity question is not ignorable if, for any score level, the proportion of ethnically identified examinees of a particular group responding correctly to the item differs from the proportion of unidentified examinees of that group responding correctly to the item. Ignorability does not depend upon the proportion of examinees from each group that identify their ethnicity. If the nonresponse is ignorable then conclusions about the DIF for a specific item are appropriate, despite missing information about group membership. Our primary purpose was to find out the extent to which the nonresponse in declared ethnicity is ignorable when using the M-H procedure.

2. DATA AND METHODS

A total of 104,330 examinees took the SAT in both May and November of 1985. These examinees were asked to provide demographic information on the SDQ for each of these two administrations, because the new form of the SDQ was pretested during that year. Some examinees provided this information consistently for both administrations (i.e., reported being a member of the same group on both occasions), some were inconsistent (i.e., changed ethnicity), some reported ethnicity at one administration and did not respond at the other, and some did not respond at either administration. We believed that this data set could provide an introductory look at the effects that ethnicity nonresponse has on the identification of DIF. We chose to examine results for the 85 item SAT-verbal test taken by all of these examinees at the May administration, because items in the May administration were given in the same order to all examinees. Items in the November administration were given to examinees in two different item orders. For this study we will consider only the approximately 94,700 examinees who identified themselves as White, Black or those who did not respond at all. The top panel of Table 1 contains a breakdown of these examinees. In addition to examinees in other ethnic groups, we omitted examinees who identified themselves as Black on one occasion and White on the other because the number of examinees who did this was so small that their omission would not affect any result in a practical way. The exact makeup of our sample is shown at the bottom of Table 1. The 94,486 examinees in our sample account for more than 91% of the total examinee sample. As is clear from the breakdown in Table 1, although the ratio of consistently identified Blacks to consistently identified examinees (2,846/72,241=3.9%) is statistically different from the ratio for the inconsistently identified examinees (686/15,656=4.4%), the practical difference between the two is small. The proximity of these two ratios was the first evidence that the inconsistently identified examinees were similar in terms of their group characteristics to the consistently identified examinees.

Insert Table 1 about here

If the consistently and inconsistently identified examinees are similar to each other, the principal question then becomes how to allocate ethnic membership among the more than six thousand examinees who never identified their ethnicity. Clearly, if these unidentified examinees were predominantly Black their performance could have a profound effect on any item statistics, because the size of the group of examinees already identified as Black is small in comparison to the White group. If they are predominately White, their effect would tend to be very modest indeed, because the size of the group of examinees already identified as White is so large.

In this study we make a series of comparisons. The comparisons differ from one another in the way the nonrespondents are classified. In each analysis, we consider Whites as the reference group and Blacks as the focal group. The analyses are summarized in Table 2. In the body of the table are the ethnic classifications used in each analysis. The column headings are the ethnic groups specified by the examinees in May and November, respectively.

Insert Table 2 about here

In the first analysis, an examinee was identified as belonging to the Black or White ethnic group if the examinee specified that ethnic group both in May and in November. Examinees were (initially) identified as nonrespondents if they didn't respond consistently to the ethnicity question on both forms of the SDQ. In subsequent stages of the analyses, members of the group of nonrespondents were included in varying ways. In the second analysis, for example, examinees who responded as Blacks or Whites at either testing time were classified in the way that they responded. Those who did not respond in either May or November were not included in the analysis. In the fifth and sixth analyses, examinees who were nonrespondents in both May and November were divided into two groups: those who scored higher on the verbal items than the median score for the total group of examinees in the study and those who scored lower than the median score for the total group of examinees. These two groups were alternately included with the Black and the White groups. For each of these analyses, items were identified for special consideration using the M-H criteria. The stratifying score was "purified" by excluding all type c items from the total score for all items other than the type c item itself. Items that were not reached by the examinee were excluded from the analyses.

Analyses 1 & 2
W-W vs. B-B
and
W-W & W-?/?-W vs. B-B & B-?/?-B

For both of these analyses, the M-H procedure "flagged" five items. Two of these items (items 5 and 16) were identified as c items and three (items 17, 50 and 63) were identified as b items (see Table 3). The consistency of the results for these two analyses indicates that examinees who identified themselves once behaved the same way on this test as did those who identified themselves consistently in both May and November. As expected, any effects of self-identification that we can see will have to be as a result of the way the 6,589 unidentified examinees are classified.

Insert Table 3 about here

Analysis 3

Same as Analysis 2 except all ?-? were considered Black

When all of the unidentified examinees were considered to be Black, this group of more than six thousand examinees swamped the 3,532 examinees who at one time or another actually identified themselves as Black. The M-H procedure did not flag **any** items in this analysis. The differences in results were not all due to sample size however. When examinees from the Black group for Analysis 3 were selected randomly so that the size of the focal group was about the same as for Analysis 2, M-H identified only one b item.

Analysis 4

Same as Analysis 2 except all ?-? were considered White

When all of the unidentified examinees were considered to be White, the M-H analysis yielded the same result that it did for Analyses 1 and 2, except that item 16 was now classified as a type b item rather than as a type c item. As can be seen in Analyses 3 and 4, the effects of nonresponse on item bias measures depends crucially on how we assign the nonrespondents. Since we have no additional information about the ethnicity of the ?-? group, the best that we can do is the sort of sensitivity studies that are de rigueur in the area of missing data (e.g. Little and Rubin, 1987; Rubin, 1987). We did many of these, assigning ethnicity to members of the ?-? group in a variety of ways, and examining the variation in the classification of items for the M-H procedure. We will report just two of these, which represent a range of possibilities that might occur.

Analysis 5

Same as Analysis 2 except:

all ?-? whose scores were above the median were considered White

and

all ?-? whose scores were below the median were considered Black

This is not the most extreme comparison possible (for example, we might have divided the ?-? population specifically to maximize DIF), but it does represent one kind of extreme. When the division in the consistent nonrespondent group occurred at the median for all examinees in the study, we found that the M-H results were the same as those in Analysis 3; there were no items that exhibited significant DIF. The explanation for the M-H result was not obvious. The reason that no items were identified may be tied to the fact that we were essentially adding a large group of individuals to the focal group whose performance on the various items (after conditioning on total score) was not, in any obvious way, tied to their ethnic designation. This somewhat random (with respect to ethnicity) assignment merely added noise to the mix, making it harder to pick out anything.

Analysis 6

Same as Analysis 2 except:

all ?-? whose scores were above the median were considered Black

and

all ?-? whose scores were below the median were considered White

This comparison is the mirror image of Analysis 5. The effect on DIF was, again, not obvious. We would have predicted, a priori, the odd mixture of examinees in the focal group would tend to diminish any DIF, as we observed in

Analysis 5. Indeed DIF was diminished, but only to a very modest degree. The M-H still flagged both items 5 and 16 (as in Analyses 1, 2 and 4), but only as the more benign type b.

3. DISCUSSION AND CONCLUSIONS

The results of the study are summarized in Table 3. Notable is the similarity of results for Analyses 1, 2 and 4 and their contrast with Analysis 3 and Analysis 5. In Analysis 6 only the two items with extreme M-H D-DIF in other analyses were identified, and those only met the type b requirements. Mean SAT-verbal scores and sample sizes for the reference and focal groups in each analysis are in Table 4.

Insert Table 4 about here

One aspect of this study deserves additional comment. Our categorization of items into classes (types b or c under the M-H guidelines) makes for gross comparisons when studying something as potentially subtle as the sensitivity of statistical measures to examinee nonresponse. For that reason, we examined the changes in DIF under assumptions about the nonrespondent groups in several ways. Results were similar whether direct differences between estimates of M-H D-DIF in the different analyses were compared or correlations between those estimates were examined. The difficulty lies in reporting these changes in a meaningful way. Because we were most interested in the effect of nonresponse on current practice, we chose to report changes in the identification of items rather than in the estimates themselves.

This study was hampered by two major obstacles. First, inasmuch as the SAT is a test with carefully scrutinized item pools, there were very few items whose performance was questionable. It would have been better, for this kind of study, had the test been a poorer one. The second obstacle was that we did not have access to the true ethnicity of even a sample of the consistently nonresponding population. A survey of these unidentified examinees might have gone a long way toward indicating a reasonable structure for imputing their ethnicity.

In conclusion, there are two pieces of good news in all of this. First, the inclusion of examinees who responded once, either in May or November, did not affect the Mantel-Haenszel results (Analysis 2 vs. Analysis 1). From a point of view taken in some of the survey sampling literature, this consistency would be an indication that the nonresponse to the ethnicity question was ignorable. Second, the M-H procedure identified no new items that were not identified when the nonresponse was treated as ignorable when the nonresponse to the question of ethnicity was treated as nonignorable. In other words, identification of the items in Analyses 3 through 6 did not involve any items that were not already identified as the result of Analyses 1 and 2.

The bad news is that, despite the survey sampling point of view described above, it is likely that consistent nonrespondents perform differently on items than one would expect based upon the performance of consistent and inconsistent respondents. Thus, the examinees whose ethnicity was never specified may contribute significantly to changes in DIF measures. Secondly, changes across analyses as a result of different assumptions about the consistent nonrespondent group indicate that, even with the gross categorizations used, differences in DIF are detected. The true ethnic categorizations for the examinees in the consistent nonresponse group can make a difference for DIF.

REFERENCES

Dorans, N. J. & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-358.

Holland, P. W. & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Little, R. J. A. & Rubin, D. B. (1987). Statistical analysis with missing data. Wiley: New York.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Phillips, A. & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. Biometrics, 43, 425-431.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley: New York.

Shepard, L. A., Camilli, G. & Williams, D. M. (1985). Validity of approximation techniques fcr detecting item bias. Journal of Educational Measurement, 22, 77-105.

Thissen, D., Steinberg, L. & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. Braun (Eds.), Test Validity (pp. 147-169). Lawrence Erlbaum Associates: Hillsdale, N.J.

Table 1

Ethnic Identification
(Number of Examinees)

		November			
		White	Black	Unspecified	Total
May	White	69,395	162	7,155	76,712
	Black	88	2,846	252	3,186
	Unspecified	7,815	434	6,589	14,838
	Total	77,298	3,442	13,996	94,736

Summary of Examinees Used in Study
(May-November)

W-W	B-B	W-?/?-W	B-?/?-B	?-?	Total
69,395	2,846	14,970	686	6,589	94,486

Table 2

Ethnic Classifications for Each Analysis
(Stratified by May-November Responses)

Analysis	W-W	B-B	W-?/?-W	B-?/?-B	?-?
1	*W*	*B*			
2	*W*	*B*	*W*	*B*	
3	*W*	*B*	*W*	*B*	*B*
4	*W*	*B*	*W*	*B*	*W*
5	*W*	*B*	*W*	*B*	*below median B/above median W*
6	*W*	*B*	*W*	*B*	*below median W/above median B*

Table 3

Items Identified by the ETS Mantel-Haenszel Guidelines

	Item Number				
Analysis	5	16	17	154	167
1	*c*	*c*	*b*	*b*	*b*
2	*c*	*c*	*b*	*b*	*b*
3					
4	*c*	*b*	*b*	*b*	*b*
5					
6	*b*	*b*			

b = Modest DIF
c = Serious DIF

Table 4

Sample Sizes
(Total N = 94,486)

Analysis	Reference Group	Focal Group	Other
1	69,395	2,846	22,245
2	84,365	3,532	6,589
3	34,365	10,121	0
4	90,954	3,532	0
5	87,568	6,918	0
6	87,751	6,735	0

Group Means

Analysis	Reference Group	Focal Group	Other
1	451	401	445
2	450	399	448
3	450	431	-
4	450	399	-
5	453	385	-
6	447	461	-