Influence of the Criterion Variable on the Identification of Differentially Functioning Test Items Using the Mantel-Haenszel Statistic

Brian E. Clauser, Kathleen Mazor, and Ronald K. Hambleton
University of Massachusetts

This study investigated the effectiveness of the Mantel-Haenszel (MH) statistic in detecting differentially functioning (DIF) test items when the internal criterion was varied. Using a dataset from a statewide administration of a life skills examination, a sample of 1,000 Anglo-American and 1,000 Native American examinee item response sets were analyzed. The MH procedure was first applied to all the items involved. The items were then categorized as belonging to one or more of four subtests based on the skills or knowledge needed to select the correct response. Each subtest was then analyzed as a separate test, using the MH procedure. Three control subtests were also established using random assignment of test items and were analyzed using the MH procedure. The results revealed that the choice of criterion, total test score versus subtest score, had a substantial influence on the classification of items as to whether or not they were differentially functioning in the American and Native American groups. Evidence for the convergence of judgmental and statistical procedures was found in the unusually high proportion of DIF items within one of the classifications and in the results of the reanalysis of this group of items. *Index terms: differential item functioning, item bias, Mantel-Haenszel statistic, test bias.*

Standardized tests have become an integral part of modern society. Test results are widely used to make decisions on acceptance and advancement in education, career advancement, and the provision of special services. Because of the increasing impact of test results in these important areas, the use of standardized tests has become a controversial social and political issue. One of the most important areas of controversy has been the issue of bias in testing. Biased test items result in one subgroup of a population having an advantage over another on the biased items and, depending on the composition of the test, sometimes on the entire test. In many cases, items that have been identified as biased favor majority group members over minority group members (Berk, 1982).

Because of the importance of the item bias issue, considerable attention has been devoted to the development and evaluation of methods for detecting differentially functioning (DIF) test items (e.g., Berk, 1982; Raju, Bode, & Larsen, 1989; Scheuneman & Bleistein, 1989). Once such items are identified, these items can be examined to determine whether the difference is due to bias in the test item or to some other cause. The Mantel-Haenszel (MH) statistic (Holland & Thayer, 1988) has emerged as one of the most popular procedures (e.g., Bennett, Rock, & Novatkoski, 1989; Mellenbergh, 1989; Zwick & Ercikan, 1989) for identifying DIF items.

The MH approach is popular for several reasons. Like the more complex conditional item bias methods based on item response theory (IRT), the MH builds on a widely accepted conceptual basis in which an item is differentially functioning if examinees of the same ability level, but belonging to different subgroups, have different probabilities of selecting the correct response. However, in contrast to some of the IRT-based item bias detection procedures, the calculations involved are relatively simple. Also,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 4, December 1991, pp. 353-359

0146-6216/91/040353-07$1.60

an acceptable test of significance for the MH exists (Holland & Thayer, 1988), and the MH procedure can produce reliable results with sample sizes considerably below the size needed for the IRT procedures (Scheuneman & Bleistein, 1989). Finally, direct comparisons of the MH and IRT procedures have demonstrated substantial agreement between them (Hambleton & Rogers, 1989).

Purpose

The use of the MH statistic, like other χ^2 and IRT procedures for detecting DIF, rests on two basic assumptions. The first assumption is that the test scores are reliable. Second, it is assumed that the test is valid, and therefore, total test score can provide a meaningful criterion against which to interpret item performance in the two groups of interest.

It is commonly assumed that the total test score is the most practical and appropriate criterion, and that its use leads to valid results. However, the stability of MH results across alternative criteria has been a concern to some researchers. For example, Hambleton, Bollwark, and Rogers (1990) compared MH results based on internal and external criteria. They found the results to be quite similar across the two standards. The present study extended this examination of the stability of the MH statistic across criteria by comparing the MH results across differing internal criteria, because tests that appear statistically to be unidimensional can contain items that measure more than one skill.

The question examined here was the extent to which total test score is an appropriate criterion against which group performance on individual items within a test can be compared. It was predicted that systematically changing the grouping of the test items analyzed would result in some items showing substantial changes in their MH statistic.

Method

Test Data and Examinee Samples

The dataset for the study was collected during the 1982 administration of the New Mexico High School Proficiency Exam (NMHSPE). It contained the responses of 23,000 students to 150 test items. The NMHSPE is intended to measure five major "life skills" areas: knowledge of community resources, consumer economics, government and law, mental and physical health, and occupational knowledge. Included in the 23,000 examinees were approximately 8,000 Anglo-American and 2,600 Native American students. These two groups were selected for comparison for three reasons: (1) their scores represented a relatively extreme case of differing ability distributions, (2) a previous study (Hambleton & Rogers, 1989) had indicated that the dataset contained several DIF items, and (3) Native American groups have been underrepresented in item bias research studies.

The Mantel-Haenszel Statistic

In calculating the MH statistic, the majority and minority groups are first matched on the criterion of interest. In most cases, the best available criterion is the total test score, either with the exclusion of the item under evaluation or with the exclusion of all items judged as having a significant likelihood of being biased. The MH statistic represents the average factor by which the odds that a reference group (majority) member gets an item correct exceeds the corresponding odds for comparable focal group (minority) members (Holland & Thayer, 1988). Holland and Thayer provided a significance test for the MH statistic that is distributed as χ^2 with 1 degree of freedom.

Computer Program

The MH computer program had been used in a previous study using this same dataset (Hambleton & Rogers, 1989). It constructs $k + 1$ score groups (where k is the total number of test items) and calculates the MH statistic for all items. It then removes all items with χ^2 values significant at the .01 level and recalculates the statistic using the remaining items as the criterion.

Procedure

Two random samples of 1,000 examinees were selected. One sample consisted of Anglo-American examinees, the other of Native American examinees. The first step in the analysis was to remove items with very low discrimination as measured by biserial correlations ($r < .1$) and that were very easy ($p \geq .90$) in the combined sample of 2,000 examinees. This was done because there seemed to be little merit in carefully analyzing test items for DIF that were contributing very little to test score variability. The result was a pool of 91 items.

Because the computer program used had a maximum of 75 items, and because previous research had raised issues regarding the stability of the MH and other item bias statistics (Hambleton & Rogers, 1989), the 91 items were divided into three groups. Items were randomly assigned to groups with the stipulation that each group contain 75 items and each item be represented in at least two of the three groups. The MH program was then run for each of the three groups separately. Items were identified as DIF only if they were identified as DIF in all of the item groups in which they appeared. This provision was made to reduce the number of items identified as DIF because of Type I errors alone.

In order to examine how the grouping of a set of test items influenced the results of the MH procedure, the items were categorized as belonging to one or more of four subsets (which became the four subtests). The MH statistics were then recalculated for items in each of the four subtests. Although the original test also was divided into subtests, examination of the items within these subtests suggested that placement of an item in a subtest had little to do with the skills required to solve an item (e.g., several of the subtests included items which depended primarily on math skills). The set of subtests used was developed based on the skills required to answer specific items correctly. The following definitions were used in constructing these four subtests:

1. Math: Any item in which computations were required to obtain the correct answer.
2. Reading: Any item in which the stem contained all the information needed to obtain the correct answer. That is, no special familiarity or prior knowledge of the item content was required to answer the item.
3. Prior Knowledge: Any item in which the stem did not contain sufficient information to allow for a correct answer; specific prior knowledge on the part of the examinee was therefore needed to answer the item correctly. Included in this category were a number of items for which there appeared to be no clearly best answer [referred to below as "No Clearly Best Answer" (NCBA)]. These items appeared to require guessing at what was intended by the item writers. Two examples of the NCBA items are shown in Figure 1.
4. Charts: Any item in which the stem was presented in other than paragraph form. This included charts, graphs, maps, tables, and price lists.

Categories 1 to 3 above were mutually exclusive. Items coded as "Charts" were in all cases coded under a second category as well. The MH computer program was run independently for each of these four subtests.

The assignment of items to categories was carried out by the first two authors independently. Initial interrater agreement exceeded .90. Consensus was reached on all items for which there were disagreements before the subtest analyses were conducted.

Finally, three additional subtests were constructed by randomly assigning each of the 91 items to one of three groups with test lengths approximately equal to those of several of the subtests formed on the basis of a skills analysis. The resulting groups of 30, 31, and 30 items each were analyzed using the MH computer program. These analyses were done to evaluate the extent to which a reduction in the number of items analyzed would be likely to influence the MH statistic. The content of these three subtests matched reasonably closely the content of the

Figure 1
Examples of NCBA Items, With Explanations of Their NCBA Rationale

129. Sid is in the National Guard and has to go to camp for two weeks every summer. When he goes to look for a job, he will be most interested in:
(A) sick leave policy
(B) retirement plan
(C) vacation policy
(D) profit-sharing plan
(The test is keyed with "C" as the correct answer. However, federal law requires employers to give members of the National Guard leave to fulfill their National Guard responsibilities. Therefore, there does not appear to be a best or correct answer.)

130. Flo needs a job that will pay about $140 a week. She will consider less if the job doesn't call for her to be away from home at night or on weekends. If she's qualified, which of the following jobs would be best for Flo?
(A) dental assistant; $125/week, 1 Saturday a month
(B) legal secretary; $130/week, Friday afternoon off
(C) weaver; $140/week, 11 p.m.-7 a.m. shift
(D) stock clerk; $120/week, 10% discount
(The test key shows "B" as the correct answer, but both "B" and "C" fulfill the stated requirements.)

total test. A concern was that the comparison of MH results with the 75-item tests and the shorter subtests would be confounded by test length. These analyses, therefore, provided an additional basis for interpreting the findings.

Results

Based on three runs of the MH computer program analyzing a total of 91 items (randomly assigned to groups of 75 with the constraint that each item be included in at least two groups), 22 items were identified as DIF in the Anglo-American and Native American groups. This result is reported in Table 1. All 22 items described in Table 1 were identified as DIF in the 75-item tests in which they appeared. Four additional items would have been identified had the criterion been identification on only one of the two runs.

These 22 DIF items were distributed across the four subtests as follows: 2 in Math, 3 in Reading, 17 in Prior Knowledge, 6 in NCBA, and 4 in Charts. (Recall that the Charts subtest overlapped with the other three subtests and the NCBA was a part of the Prior Knowledge subtest.)

When these four subtests were analyzed, a number of changes in the MH results were observed. In fact, approximately a third of the test items (7 of 22) ceased to be DIF when analyzed within the subtests (1 of the 2 Math, 1 of the 3 Reading, 5 of the 17 Prior Knowledge, and 3 of the 4 Chart items). In general, these results confirmed the original hypothesis that changes in item grouping will change the MH results.

In an effort to better understand these results, the three randomly selected control subtests described above were analyzed. This provided subtests similar in numbers of items to the subtests described above, but without constraints on the skills measured in the subtests. The result of the analysis of control subtest 1 was that none of the items changed their DIF status. Only minor changes in the DIF of items were noted in the other two subtests as well. The complete results appear in Table 1.

The 12 items classified as NCBA provided some support for the validity of the main results. These items were assumed to be truly flawed—by definition there was no clearly best answer provided. A "guess" as to what the item writer intended seemed to be required in these items, and it is reasonable to expect that Native American examinees would be at a disadvantage compared to majority group examinees. Six of the 12 NCBA items were identified as DIF in the two groups in the 75-item test and remained so in the subtest analyses. These flawed test items were consistently identified regardless of the subtest in which they were located. This result contrasts sharply

Table 1
Number of Items, Number of DIF Items in the 75-Item Test (DIF-75), Number of Items That Were No Longer DIF When Analyzed Within a Subtest (No Longer DIF), and Number of New Items Identified as DIF When Analyzed Within a Subtest (New Items DIF)

Category and Total*	Subtest							
			Prior Knowledge			Control Subtest		
	Math	Reading	All	NCBA	Charts	1	2	3
No. of Items								
91	27	15	49	12	19	30	31	30
DIF-75								
22	2	3	17	6	4	7	7	8
No Longer DIF								
7	1	1	5	0	3	0	1	3
New Items DIF								
11	2	3	5	2	4	2	5	5

*Because NCBA and Charts were not mutually exclusive of the other three subtest categories, results in the Total column cannot be obtained by summing across subtest results.

with the results of the other 37 items in this subtest for which there clearly was a best answer. Eleven of these 37 items were identified as DIF in the 75-item test. Five of these items were not so labeled in the subtest. The shift in results was 45% (5 of 11). A complete analysis of the comparison of DIF and Non-DIF items in the 75-item test and the various subtests of interest appears in Table 2.

A less easily predicted phenomenon was also noted as a result of the subgroup runs (see Table 1). In each case, some items not previously identified as DIF were classified as such by these runs: 2 additional Math items, 3 additional Reading items, 5 additional Prior Knowledge items (including two in the NCBA group), and 4 additional Charts items were identified. When only the mutually exclusive subtests were considered, allowing each of the 91 items to be analyzed only once, 10 new DIF items were identified—from a total of 69 not previously identified. This is well beyond the number that might be expected due to chance alone using the .01 significance level.

Discussion

The results suggest that test developers using the MH statistic to assess item bias in tests should be cautious in interpreting the results. When the original set of items analyzed here was regrouped and reassessed within separate subtests, 32% of the DIF items (7 of 22) were no longer found to be DIF. This represents 8% of the total item pool and is well beyond what would be expected as a typical false positive error rate using a .01 significance level. To what extent the substantially differing ability distributions for the majority and minority samples exaggerated this effect is an empirical question, but it may be prudent for test developers to be especially cautious when such differences exist. It appears that the context (test) in which items are studied can (and will) influence the results. Practitioners might be encouraged, for example, when conducting large item calibration and DIF studies, to consider groupings of test items as one of the variables in preparing a test design.

The data also revealed an unpredicted and less easily explained phenomenon. When items were grouped into subtests and then reanalyzed within these subtests, the percentage of DIF items increased. This result was consistent across all subtests, including the three "control" subtests. The control subtests differed from the original test only in number of items. Thus, in general, as the length of the test decreased, the number of additional items identified as DIF increased. Again, this result might have serious implications for test developers using the MH statistic. Specific

Table 2
Comparison of DIF and Non-DIF Items in the Total Test and Subtests

Subtest and Item Type	Subtest		Level of Agreement
	DIF	Non-DIF	
Math, 27 Items			
DIF	1	1	88.9%
Non-DIF	2	23	
Reading, 15 Items			
DIF	2	1	73.3%
Non-DIF	3	9	
Prior Knowledge, 49 Items			
DIF	12	5	79.5%
Non-DIF	5	27	
Prior Knowledge, 37 Items (Factual Non-DIF)			
DIF	6	5	78.3%
Non-DIF	3	23	
Knowledge, 12 Items (NCBA Non-DIF)			
DIF	6	0	83.3%
Non-DIF	2	4	
Charts, 19 Items			
DIF	1	3	63.2%
Non-DIF	4	11	
Control 1, 30 Items			
DIF	7	0	93.3%
Non-DIF	2	21	
Control 2, 31 Items			
DIF	6	1	80.6%
Non-DIF	5	19	
Control 3, 30 Items			
DIF	5	3	73.3%
Non-DIF	5	17	

recommendations for test developers in this area must await further research.

One final finding of the present study is the apparent usefulness of the NCBA subcategory within the Prior Knowledge subtest. 50% of the items classified as NCBA were shown to be DIF. Although it is assumed that most test developers routinely exclude all items fitting this definition, the presence of such items in this test provides clear evidence of convergence between the statistical and judgmental approaches to identifying bias. In the case of these items, not only is the classification of NCBA a good indicator of the likelihood of statistically demonstrable bias, but all of the six items in this category originally identified as DIF remained in that status when reanalyzed as part of the Prior Knowledge subtest. This kind of convergence would seem to argue for validity of both the statistical and judgmental techniques employed.

Although these results should provide some cautions to test developers using the MH statistic and suggest the need for additional study, the nature of the data analyzed provides little basis for a theoretical explanation of the findings. It is, for example, impossible to make meaningful statements regarding the relative accuracy of the alternate MH analyses. It could be argued that items showing a change in status across analyses (i.e., across criteria) should be considered non-biased. This would be conservative with respect to Type I error. However, for the argument to stand, it is necessary to know the power characteristics of the MH statistic, particularly for varying test lengths. But the power characteristics of the MH statistic are not known when it is applied in an iterative way as it was here.

Another issue raised, but not resolved by these data, was the impact of dimensionality on the MH statistic. Although there are several indices of dimensionality in common use, none have been shown to be dependable (Hattie, 1985). The results obtained here may be explained at least in part by changes in dimensionality of the regrouped tests. Initial analysis examining indices such as the ratio of the first to the second eigenvalue (Ackerman, 1989) did not show clear support for a change in dimensionality. However, these indices may be confounded by test length. As with the power issue referred to above, the part dimensionality plays in these results may best be examined with a simulation study so that the level of differential functioning as well as the dimensionality can be specified.

These data raise important questions regarding the functioning of the MH statistic. Additional research is needed to determine the extent to which these results can be generalized to other test data and to provide a theoretical explanation of the results. The data presented here strongly suggest that the phenomena observed represent systematic rather than random occurrences. To the extent that these results are generalizable to

other datasets, they may lead to the development of important guidelines for test developers in using the MH statistic.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113–127.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille Edition. *Journal of Educational Measurement, 26,* 67–79.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore MD: Johns Hopkins University Press.

Hambleton, R. K., Bollwark, J., & Rogers, H. J. (1990). *Factors affecting the stability of the Mantel-Haenszel item bias statistic* (Rep. No. 203). Amherst: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluative Research.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313–334.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

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

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2,* 1–13.

Scheuneman, J., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2,* 255–275.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*(1), 55–66.

Acknowledgments

The authors are grateful to John Hattie, University of Western Australia, for helpful comments on the interpretation of results. This project was supported by a research grant from the School of Education at the University of Massachusetts at Amherst. An earlier version of this paper was presented at the 1990 annual meeting of the American Educational Research Association, Boston MA, U.S.A.

Author's Address

Send requests for reprints or further information to Ronald K. Hambleton, Laboratory of Psychometric and Evaluative Research, School of Education, University of Massachusetts, Amherst MA 01003, U.S.A.