ED 359 190 TM 019 778

| | |
|---|---|
| AUTHOR | Mazor, Kathleen M.; And Others |
| TITLE | Identification of Non-Uniform Differential Item Functioning Using a Variation of the Mantel-Haenszel Procedure. |
| PUB DATE | Mar 93 |
| NOTE | 15p.; Information f. m this article was presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 13-15, 1992). |
| PUB TYPE | Reports - Evaluative/Feasibility (142) -- Speeches/Conference Pap-rs (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | Comparative Testing; *Computer Simulation; *Item Bias; Item Response Theory; Research Methodology; *Statistical Distributions; *Test Items |
| IDENTIFIERS | *Mantel Haenszel Procedure |

ABSTRACT

The Mantel-Haenszel (MH) procedure has become one of the most popular procedures for detecting differential item functioning (DIF). One of the most troublesome criticisms of this procedure is that while detection rates for uniform DIF are very good, the procedure is not sensitive to non-uniform DIF. In this study, examinee responses were generated to simulate both uniform and non-uniform DIF. Responses for 3 groups of 1,000 examinees (1 reference group and 2 focal groups) were simulated using a 3-parameter logistic item response theory model. A standard MH procedure was used first. Then examinees were split into two samples, by breaking the full sample at approximately the middle of the test score distribution. The tests (25 simulated tests) were then re-analyzed, first with the low performing sample, and then with the high performing sample. This variation improved detection rates of non-uniform DIF considerably over the total sample procedure without increasing the Type I error rate. Items with the largest differences in discrimination and difficulty parameters were most likely to be identified. Two tables present study data. (Author/SLD)

# IDENTIFICATION OF NON-UNIFORM DIFFERENTIAL ITEM FUNCTIONING USING A VARIATION OF THE MANTEL-HAENSZEL PROCEDURE

Kathleen M. Mazor, Brian E. Clauser, Ronald K. Hambleton
University of Massachusetts at Amherst

## Abstract

The Mantel-Haenszel (MH) procedure has become one of the most popular procedures for detecting differential item functioning (dif). One of the most troublesome criticisms of this procedure is that while detection rates for uniform dif are very good, the procedure is not sensitive to non-uniform dif. In this study, examinee responses were generated to simulate both uniform and non-uniform dif. A standard MH procedure was used first. Then examinees were split into two samples, by breaking the full sample at approximately the middle of the test score distribution. The tests were then re-analyzed, first with the low performing sample and then with the high performing sample. This variation improved detection rates of non-uniform dif considerably over the total sample procedure without increasing the Type I error rate. Items with the largest differences in discrimination and difficulty parameters were most likely to be identified.

Lab Report 227

# IDENTIFICATION OF NON-UNIFORM DIFFERENTIAL ITEM FUNCTIONING USING A VARIATION OF THE MANTEL-HAENSZEL PROCEDURE

Kathleen M. Mazor, Brian E. Clauser, Ronald K. Hambleton
University of Massachusetts at Amherst

The issue of bias in testing is an important one, with serious legal, ethical and social ramifications. In recent years, research in the area of item bias, or differential item functioning (*dif*) has proliferated. *Dif* is said to exist when examinees *of the same ability* but from different groups have differing probabilities of success on a given item. While there is wide agreement on the appropriateness of this definition of *dif*, there is less agreement as to which statistical procedures are best suited for detecting it.

One of the most popular procedures currently in use is the Mantel-Haenszel (MH) procedure (Holland and Thayer, 1988). There are a number of advantages typically attributed to this procedure, among them being that it is theoretically defensible, has an associated test of statistical significance, can be used with fewer examinees than are required for procedures based on item response theory, is easy to program, and is relatively inexpensive in terms of computer time. Empirical studies of the MH procedure under a number of different conditions using both real and simulated data have yielded generally positive results (Hambleton and Rogers, 1989; Mazor, Clauser and Hambleton, 1992; Rogers, 1989). There is one caveat however that has been issued repeatedly: the MH procedure is not sensitive to non-uniform *dif* (Scheuneman and Bleistein, 1989; Swaminathan and Rogers, 1990). Non-uniform *dif* may be said to be present when the difference in probability of success between two groups is not constant across ability levels. That is, there is an interaction between group membership and ability. In item response theory

(IRT), this interaction may be thought of as a difference in the item discrimination (a-) parameters.

Although many assume that non-uniform dif is relatively rare in practice, studies which have used techniques sensitive to non-uniform dif have found it to be present in real data sets (e.g. Be..nett, Rock, and Kaplan, 1987; Ellis, 1989; Hambleton and Rogers, 1989; Linn, Levine, Hastings, and Wardrop, 1981; Mellenbergh, 1982). Thus, insensitivity to this type of dif is a real concern. In fact, it is probably the primary drawback to the MH procedure at present. Yet, although it is widely acknowledged to be an issue, this area is one which has not been fully researched.

A recent study by Rogers (1989) suggests that the generally accepted assertion that the MH procedure is insensitive to non-uniform dif may be an overgeneralization. Rogers (1989) compared the MH procedure to a logistic regression procedure and found that although the logistic regression procedure was superior in detecting non-uniform dif, the MH procedure did identify a number of these items. Items which were easy or difficult (and displayed non-uniform dif) were fairly consistently identified. The items which the MH procedure was most likely to miss were non-uniform dif items of medium difficulty. For these items, the item characteristic curves (ICCs) crossed close to the middle of the ability range, and the differences between the two groups essentially canceled each other out. This finding is not urprising, as the MH procedure produces a signed statistic, and negative differences in one part of the score range can offset positive differences elsewhere. Because the MH procedure is also weighted by the number of examinees at each ability level, it can also be predicted that the extent to which this offsetting occurs will be a function of both the item difficulty and the examinee ability distributions. This prediction suggests that if the examinees are split into two samples, such that there is a high performing sample and a low performing sample and if the MH procedure is run for each sample separately, then it may be able to identify items showing non-uniform *dif*.

The purpose of the present study was to ascertain whether a simple modification of the MH procedure (re-analyzing the data separately for high and low performing groups) can improve detection rates for items showing non-uniform *dif*. This variation is a simple, easily implemented procedure, and if effective would overcome one of the major deficits typically attributed to the MH procedure. Although IRT-based methods are sensitive to non-uniform *dif*, these methods are usually complex and expensive, and sample size and computing requirements may be prohibitive in many practical settings. Logistic regression, which has also been demonstrated to be effective in identifying non-uniform *dif* (Swaminathan and Rogers, 1990), is less complex and expensive than IRT-based methods, but more so than the Mu procedure. In addition, the recent profusion of research with the MH procedure has made it familiar to many practitioners, most of whom are not yet familiar with logistic regression procedures. In the final phase of this study an examination is made of parameters of the items missed by the traditional MH procedure and those missed by the proposed variation to determine whether certain combinations of item parameters are more likely to be missed than others.

## Method

Responses for three groups of 1000 examinees each were simulated by using a three-parameter logistic IRT model. One reference group was simulated, with ability normally distributed with a mean of 0 and a standard deviation of 1. Two focal groups were simulated: (a) the first with an ability distribution identical to that of the reference group and (b) the second, with a mean one standard deviation below the reference group mean.

Twenty-five tests were simulated, with each test containing 59 non-dif items, and 16 studied items, for a total of 75 items. All c's were set to be .20. The a and b parameters for the 59 non-dif items were selected randomly from published item statistics for a recent edition of the Graduate Management Admissions Test (Kingston, Leary, and Wightman, 1988).

Four hundred studied items were generated and then assigned to one of the 25 tests. One set was attached to each of the 25 tests which was simulated. For these 400 items, five levels of difficulty were chosen with the b's for the reference group being set at -1.5, -1.0, 0, 1.0, or 1.5. Four levels of b differences between the reference and focal groups were chosen, so that the b's for the focal groups were higher by 0, .30, .60, or 1.00. Four levels of discrimination were chosen such that a was equal to .25, .60, .90 or 1.25. Five levels of differences between the a's were chosen, with these differences set to be 0, .25, .50, .75 and 1.0. All of these conditions were completely crossed (five levels of item difficulty x four levels of b value difference between the reference and focal groups x four levels of item discrimination x five levels of a value difference between the reference and focal groups). The resulting 400 items were then randomly grouped into blocks of 16 items and added to the core test of 59 items. In total, 25 tests were needed to study the 400 dif items (16 items per test).

An MH procedure program written by Rogers and Hambleton (1989) was used to analyze each test. First, each test was analyzed by using all examinees and using total test score as the matching criterion. Items exhibiting dif were identified. The mean observed score of the entire sample (reference and focal groups combined) was calculated, and the sample was split at this score, so that those scoring less than the mean were assigned to the low performing sample, and those scoring above the mean were in the high performing sample.

The MH procedure was then repeated, but this time using only the low performing sample in the analysis. Again, items exhibiting dif were identified. Finally, the MH procedure was run a third time -- this time with only the high performing sample in the analysis. Thus the MH procedure was run three times, once with the total sample, once with only those examinees in the lower half of the score distribution, and again with only those in the upper half. A simple modification to the original MH program made it possible for one implementation to accomplish all three runs. Identification rates for each procedure and the three procedures together were examined. In all cases an item was considered identified as exhibiting dif on a given run if the MH procedure chi-square statistic for that item was significant at the .01 level.

## Results

The first important result of this study is that the standard MH procedure was able tc identify many of the items which had been generated to simulate non-uniform dif. As can be seen from Table 1, more than 60% of the 320 items with non-uniform dif were identified using the MH procedure on the full sample. This result was true with reference and focal groups with equal and unequal ability distributions.

When examinees from equal ability distributions were compared, the examinees from groups simulated to have equal ability distributions including the two additional split-sample MH runs resulted in identification of 44% of the non-uniform dif which had been missed by the standard (total sample) analysis. When the criterion for an item to be classified as dif was that it be so classified on at least one of the three MH analyses, 82% of the non-uniform dif items were correctly classified. Of the 20 studied items which did not concain dif, only one item was incorrectly classified, and only on the run using the total sample. Thus, it appears the Type I error rate was not inflated.

When examinees from unequal ability distributions were used, 61% of the dif items were identified by employing the total sample analysis. Approximately 37% of the items missed with the total sample analysis were identified when the analysis was repeated on the two halves of the test score distribution, separately. A total of 76% of the dif items were identified on at least one of the three analyses. Again, of the 20 studied items which did not contain dif, only one false positive error was observed.

The detection rates with respect to the item parameters are presented in Table 2 for the equal ability group comparisons. Because the trends for both equal and unequal ability distributions were similar, only results for equal ability distribution comparisons are presented. From this table, it can be seen that as the between group differences of the as increased, detection rates increased. As between group differences in the bs increased, detection rates increased also. In fact, for b-differences of .6 or 1.0 very few items were missed regardless of the a-difference, or the value of a for the two groups. This pattern of results was true for the unequal ability comparison as well, although overall detection rates were somewhat lower.

The split sample runs were able to pick up a number of items that had been missed on the full run (as indicated in Tables 1 and 2). Again, larger a-differences ($\geq .5$) and larger b-differences ($\geq .3$) were associated with higher detection rates. When equal ability groups were compared, there was a slight tendency for the split sample runs to identify items with bs of -1., 0, or 1.0 over the more extreme items. In comparisons of groups of unequal ability, there seemed to be a tendency for easier items to be identified, and the most difficult items to be missed. The more 'iscriminating items were slightly less likely to be flagged, although this result seemed to depend to some extent on the difficulty of the item as well. This may result from the fact that as the a-parameter increases, the area between the curves associated with a given between group difference in a parameters decreases.

## Discussion

The results of this study suggest that the MH procedure is able to identify a relatively high proportion of items with non-uniform dif. That the standard (total sample) MH procedure identified a substantial number of non-uniform dif items could have been anticipated because, when the ICCs cross at either the low or the high end of the ability range, the differences between the groups being predominantly of the same sign will not cancel each other out. However, this finding does call into question the common assertion that the MH procedure is insensitive to non-uniform dif and suggests that this lack of sensitivity is true only in some circumstances.

In conclusion, the results have indicated that a simple modification to the standard MH procedure increased detection rates with respect to non-uniform dif items. By dividing tne total sample into a high performing subsample and a low performing subsample, and then re-running the MH procedure on each sample separately, it was possible to increase identification rates substantially without increasing the Type I error rate. This variation on the standard MH procedure is simple and carries no apparent risks. Additionally, may allow practitioners to identify non-uniform dif items which may be missed by the standard MH procedure within a framework that is likely to already be part of their dif screening protocol. While there is no reason to suggest that this procedure is to be theoretically favored over IRT procedures or logistic regression, it provides a simple alternative which may carry many of the practical advantages of these procedures.

Additional research with this variation is necessary to test its usefulness in actual test situations. A simultaneous comparison of the results of this procedure with those of a three-parameter IRT <u>dif</u> analysis, and those of a logistic regression analysis would also be useful.

## References

Bennett, R. E., Rock, D. A., and Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. Journal of Educational Measurement, 24(1), 56-64.

Ellis, B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74(6), 912-921.

Hambleton, R. K., and Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

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

Kingston, N., Leary, L., and Wightman, L. (1988). An exploratory study of the applicability of item response theory to the Graduate Management Admission Test (GMAT Occasional Papers). Princeton, NJ: Graduate Management Admission Council.

Linn, R. L., Levine, M. V., Hastings, C. N., and Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Mazor, K. M., Clauser, B. E. and Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52(2), 443-451.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7(2), 105-118.

Rogers, H. J. (1989). Item bias investigation with logistic regression. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Rogers, H. J., and Hambleton, R. K. (1989). MH: A Fortran V program to compute the Mantel-Haenszel statistic for detecting differential item functioning. Amherst, MA: University of Massachusetts, School of Education.

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

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

January 8, 1993

Table 1

Detection Rates of *Non-Uniform* *dif* Items
(n=320)

|  | Ability Distributions |  |
|---|---|---|
| Group(s) | Equal | Unequal |
| Full Sample | 68% | 61% |
| Full Sample, or<br>Low Ability, or<br>High Ability | 82% | 76% |

Table 2

Items Identified as Differentially Functioning
(Equal Ability Distributions)

| a difference | b difference | b=-1.5 a=.25 | .60 | .90 | 1.25 | b=-1.0 a=.25 | .60 | .90 | 1.25 | b=0.0 a=.25 | .60 | .90 | 1.25 | b=1.0 a=.25 | .60 | .90 | 1.25 | b=1.5 a=.25 | .60 | .90 | 1.25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | X | X | X | X | / | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| 1.0 | 0.6 | X | / |  | X | / | / | X | X | X | X | X | X | / | X | / | X | X | X | X | X |
| 1.0 | 0.3 | X | X | * |  | X | X | / | / | / | X | / | X | X | X | X |  | X | X | X |  |
| 1.0 | 0.0 | X | X | X | X | X | X | X | X | / | / | / | / | X | X | / |  | X | * | / |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| .75 | 1.0 | / | X | X | X | / | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| .75 | 0.6 | X | X | X | / | / | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| .75 | 0.3 | X | X | X |  | X | / |  | / | / | / | X |  | X | X | X | / | X | X | X | / |
| .75 | 0.0 | X | X | X |  | X | X | X | * | / | / |  | / | X |  |  |  | X | X | X | / |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| .50 | 1.0 | / | X | X | X | X | / | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| .50 | 0.6 | X |  | X | X | / | / | X | X | X | X | X | X | X | X | X | * | X | X | X | X |
| .50 | 0.3 | X | X |  |  | / | * |  |  | / | / | X | / | X | X |  |  | X | X | * |  |
| .50 | 0.0 | X | X | X | X | X | X | X | * | / | / |  |  | X | / |  | / | X | X |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| .25 | 1.0 |  | X | X | X |  | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| .25 | 0.5 |  |  | X | X |  | X | X | X | X | X | X | X | X | X | X | X | X | * | X | / |
| .25 | 0.3 | X |  |  |  |  | / |  | X | / |  |  | X | X | X | * |  | X |  |  |  |
| .25 | 0.0 | X | X |  |  | X | X | X |  |  | * |  |  | X |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0 | 1.0 | X | X | / | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| 0 | 0.6 | / | X | X | X |  | X | X | X |  |  | X | X |  | X |  | X |  | X |  | / |
| 0 | 0.3 |  |  |  |  | / | * |  |  |  |  |  | X |  |  |  | / |  |  |  |  |
| 0 | 0.0 |  |  |  |  | * |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

KEY: X = flagged by total and split samples.
/ = flagged by split samples only.
* = flagged by total sample only.

Lab Report 227

## End Notes

This paper is a reproduction of several portions of the document Laboratory of Psychometric and Evaluative Research Report No. 227. Amherst, MA: University of Massachusetts, School of Education. Information contained within this article was reported in a paper presented at the meeting of the National Council on Measurement in Education, San Francisco, 1992.

Brian E. Clauser is presently with the National Board of Medical Examiners, Philadelphia, PA.