DOCUMENT RESUME

ED 409 365 TM 026 903

AUTHOR Kwak, Nohoon; And Others
TITLE An Unsigned Mantel-Haenszel Statistic for Detecting Uniform and Nonuniform DIF.
PUB DATE 27 Mar 97
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chi Square; Identification; *Item Bias; *Test Items
IDENTIFIERS *Mantel Haenszel Procedure; Mean (Statistics)

ABSTRACT

This paper introduces a new method for detecting differential item functioning (DIF), the unsigned Mantel-Haenszel (UMH) statistic, and compares this method with two other chi-square methods, the Mantel-Haenszel (MH) and the absolute mean deviation (AMD) statistics, in terms of power and agreement between expected and actual false positive rates. Three hundred datasets included items with uniform DIF; another 300 datasets included items with nonuniform DIF; and the other 300 datasets included items with both uniform and nonuniform DIF. All methods produced higher false positive rates than the theoretically expected false positive rates after application of a purification procedure. The second step of the purification procedure produced more false positives for the MH and UMH methods than the first step, but it reduced false positives for the AMD method. The two-step purification procedure also reduced power in most conditions for all three methods. (Contains 11 tables and 30 references.) (SLD)

AERA Annual Meeting, Session 38.38, Thursday March 27th, 1997



# AN UNSIGNED MANTEL-HAENSZEL STATISTIC FOR DETECTING UNIFORM AND NONUNIFORM DIF

Nohoon Kwak
University of Minnesota

Mark L. Davison
University of Minnesota

Ernest C. Davenport, Jr.
University of Minnesota

Correspondence concerning this manuscript should be sent to Nohoon Kwak, Department of Educational Psychology, University of Minnesota, 178 Pillsbury Dr. S. E., Minneapolis, MN 55455

## Abstract

This paper introduces a new method for detecting differential item functioning (DIF), the unsigned Mantel-Haenszel (UMH) statistic and compares this method with two other chi-square methods, the Mantel-Haenszel (MH) and the absolute mean deviation (AMD) statistics, in terms of power and agreement between expected and actual false positive rates. Three hundred datasets included items with uniform DIF; another three hundred datasets included items with nonuniform DIF; and the other three hundred datasets included items with both uniform and nonuniform DIF. All methods produced higher false positive rates than the theoretically expected false positive rates after application of a purification procedure. The second step of the purification procedure produced more false positives for the MH and the UMH methods than the first step but it reduced false positives for the AMD method. Additionally, the two-step purification procedure reduced power in most conditions for all three methods.


Key words: Unsigned Mantel-Haenszel statistic, absolute mean deviation statistic, Mantel-Haenszel statistic, differential item functioning (DIF), uniform DIF, nonuniform DIF.

# AN UNSIGNED MANTEL-HAENSZEL STATISTIC FOR DETECTING UNIFORM AND NONUNIFORM DIF

Nohoon Kwak, Mark L. Davison, and Ernest C. Davenport, Jr.

University of Minnesota

Differential item functioning (DIF) has been an important issue in educational and psychological measurement since the 1960's. DIF exists if equally able individuals from different groups have different probabilities of answering an item correctly (Holland & Thayer, 1988; Shepard, Camilli, & Averill, 1981). Generally, it can be defined as follows;

$$p(X = 1 \mid g, \theta) \neq p(X = 1 \mid \theta),$$

where $X$, $g$, and $\theta$ express a dichotomous response, a group membership, and an ability level, respectively (Mellenbergh, 1989; Millsap & Everson, 1993). There are two kinds of DIF, uniform DIF and nonuniform DIF (Mellenbergh, 1982). Uniform DIF occurs when the probabilities of success on the item for one group are consistently higher than those for the other group over all trait levels. In contrast, nonuniform DIF occurs when there is an interaction between trait level and group membership.

Since Holland and Thayer (1988) introduced the Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959) for detecting DIF, it has been one of the most popular methods because of its computational simplicity, ease of implementation, and usability for small samples. However, several researchers (Hambleton & Rogers, 1989; Holland & Thayer, 1988; Kwak, 1994; Kwak, Davenport & Davison, 1997; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) found that although the MH statistic is sensitive to uniform DIF, it is relatively insensitive to nonuniform DIF. Methods for detecting DIF should be able to detect items with both nonuniform and uniform DIF because both appear in empirical studies (Bennett, Rock, & Kaplan, 1987; Ellis, 1989; Hambleton & Rogers, 1989; Linn, Levine, Hastings, & Wardrop, 1981; Mellenbergh, 1983).

There are several methods for identifying items that exhibit DIF but few of them identify both uniform and nonuniform DIF. Although the logistic regression (LR) method (Rogers & Swaminathan; 1993; Swaminathan & Rogers, 1990), the full chi-square statistic (Bishop, Fienberg, & Holland, 1975), and the cross simultaneous item bias (CSIB) test (Li & Stout, 1996) can detect both uniform and nonuniform DIF, these methods have at least one flaw. One major problem of the LR

procedure and the full chi-square statistic is that the test statistic, $G^2$ or $\chi^2$, is not distributed as chi-square when the cell frequencies of tables are sparse (Agresti, 1990, 1996) and this problem may be more serious for multiple degree of freedom tests such as the LR and the full chi-square methods than 1 degree of freedom tests such as the AMD or the MH method. Another limitation of the LR method is lower power for detecting uniform DIF than detecting nonuniform DIF because the test statistic is based on a chi-square distribution with two degrees of freedom rather than one (Clauser, Nungster, Mazor, & Ripkey, 1996; Swaminathan & Rogers, 1990). Finally, model fitting of the LR method is slower and more expensive to implement than the MH procedure, particularly when combined with iterative purification procedures (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). A major problem of the CSIB test is that it requires a "valid subtest", because it is hard to constitute a valid subtest before checking whether or not a test contains items with DIF.

Since Holland & Thayer (1988) suggested a two-step purification procedure, many studies have used the two-step purification procedure but only two studies (Kwak, Davison, & Davenport, 1997; Miller & Oshima, 1992) evaluated it systematically. Miller and Oshima (1992)

indicated that the purification procedure reduced the false positives for the Mantel-Haenszel statistic. Kwak, et al. (1997) found that the two-step purification procedure reduced the false positives or increased power for the MH, the full chi-square, and the absolute mean deviation (AMD: Kwak, Davison, & Davenport, 1997) methods. However, these studies used approximately the same ability groups. Therefore, their research on the effectiveness of purification should be extended to groups with different ability distributions.

The primary goal of the proposed paper is to introduce a new statistic called the unsigned Mantel-Haenszel (UMH) procedure for detecting both uniform and nonuniform DIF and to compare this method with the MH and the absolute mean deviation (AMD) statistics. The secondary goal is to evaluate the effect of the two-step purification procedure for the MH, the AMD, and the UMH methods.

## DIF Detection Methods

### Unsigned Mantel-Haenszel (UMH) Statistic

One drawback of the MH statistic is that it is not sensitive to nonuniform DIF (Kwak, 1994; Kwak, Davenport, & Davison, 1997; Rogers & Swaminathan, 1993; Swaminathan, & Rogers, 1990). However,

the proposed UMH statistic, which is a modification of the original MH statistic, should be able to detect nonuniform DIF as well as uniform DIF. This statistic assumes no three factor interaction. For this statistic, the "reference" group is always the group which has the higher proportion-correct (H) on a studied item in score group $j$, and the "focal" group is always the group which has the lower proportion-correct (L) on the studied item in score group $j$. The UMH statistic is based on Table 1.

---

Insert Table 1 Here

---

In Table 1, $H_{1j}$ is the number of examinees who answered an item correctly in score group $j$ and $H_{0j}$ is the number of examinees who answered the item incorrectly in score group $j$ for the higher proportion-correct group (i.e., the reference group). $L_{1j}$ and $L_{0j}$ are similarly defined for the lower proportion-correct group (i.e., the focal group) in score group $j$. $N_{Hj}$, $N_{Lj}$, $N_{1j}$ and $N_{0j}$ are the marginal totals. $N_j$ is the total number of examinees in the $j$th score group. Each group, $j$, is conditioned on total score and there will be as many tables as there are different score groups.

Given the totals of the $2 \times 2$ tables, cell counts from different $2 \times 2$ tables are independent. Thus, $\Sigma_j H_{1j}$ has expectation, $\Sigma_j E(H_{1j})$, with

corresponding variance, $\Sigma_j \mathrm{Var}(H_{1j})$. Thus, the UMH statistic for detecting nonuniform DIF as well as uniform DIF is

$$\chi^2_{UMH} = \frac{\left[\left|\sum_{j=1}^{J} H_{1j} - \sum_{j=1}^{J} E(H_{1j})\right| - 0.5\right]^2}{\sum_{j=1}^{J} Var(H_{1j})} \tag{1}$$

where

$$\mathrm{E}(H_{1j}) = \frac{N_{Hj} N_{1j}}{N_j} \tag{2}$$

and

$$\mathrm{Var}(H_{1j}) = \frac{N_{1j} N_{0j} N_{Hj} N_{Lj}}{N_j^2 (N_j - 1)}. \tag{3}$$

Under the null hypothesis of conditional independence, the UMH statistic has an asymptotic chi-square distribution with $df = 1$. The UMH chi-square formula indicates that the UMH statistic equals the MH statistic when uniform DIF exists, and the UMH statistic is always larger than the MH statistic because the quantity "$H_{1j}$ - E($H_{1j}$)" is always positive in all score groups even when nonuniform DIF exists. Hence, this statistic can detect nonuniform DIF as well as uniform DIF.

The Mantel-Haenszel and the Absolute Mean Deviation Statistics

The MH and the AMD statistics were selected for comparing the UMH statistic because the MH method is a more powerful technique for

detecting uniform DIF and the AMD method is a more powerful technique for detecting nonuniform DIF (Kwak, Davison, & Davenport, 1997). These statistics are based on Table 2.

---

Insert Table 2 Here

---

The formula for the MH is

$$\chi^2_{MH} = \frac{\left[\left|\sum_{j=1}^{J} A_j - \sum_{j=1}^{J} E(A_j)\right| - 1/2\right]^2}{\sum_{j=1}^{J} Var(A_j)}. \tag{4}$$

where

$$\mathrm{E}(A_{1j}) = \frac{N_{Rj} N_{1j}}{N_j} \tag{5}$$

and

$$\mathrm{Var}(A_{1j}) = \frac{N_{1j} N_{0j} N_{Rj} N_{Fj}}{N_j^{\,2}(N_j - 1)}. \tag{6}$$

The formula for the AMD statistic (Kwak, Davenport, & Davison, 1997) is

$$\chi^2_{AMD} = \frac{\left[\sum_{j=1}^{J}\left|A_j - E(A_j)\right| - \sum_{j=1}^{J} E\left|A_j - E(A_j)\right|\right]^2}{\left[\sum_{j=1}^{J} Var\left|A_j - E(A_j)\right|\right]} \tag{7}$$

where

$$E[|A_j - E(A_j)|] = \left[\frac{2N_{Rj}N_{Fj}N_{1j}N_{0j}}{\pi N_j^3}\right]^{1/2} \tag{8}$$

and

$$Var[|A_j - E(A_j)|] = \frac{(\pi - 2)N_{Rj}N_{Fj}N_{1j}N_{0j}}{\pi N_j^3}. \tag{9}$$

Both the MH and the AMD statistics have one degree of freedom.

## Methods

The current study used 900 simulated data sets based on the three-parameter logistic model to compare a new method, the unsigned Mantel-Haenszel (UMH) statistic, with other chi-square statistics (MH and AMD) for detecting DIF. Three hundred data sets include items with uniform DIF. Another three hundred data sets include items with nonuniform DIF. The other three hundred data sets include items with both uniform and nonuniform DIF. In this simulation, we compared the performance of the measures when the null hypothesis $H_0$: $p_{Rj} = p_{Fj}$ is true to examine the agreement of actual and expected false positive (FP) rates. Further, we compared the power of the measures in three kinds of data sets: (1) the first data sets included items with uniform DIF caused

solely by the difference of difficulty parameters for two ICCs, (2) the second data set included items with nonuniform DIF caused by the difference of discrimination parameters (i.e., symmetric nonuniform DIF) or both item difficulty and discrimination parameters (i.e. nonsymmetric nonuniform DIF) for two ICCs, and (3) the third data set included items with both uniform and nonuniform DIF.

Data Generation

The simulation of item response data for examinees was based on the three-parameter logistic IRT model. The three-parameter model (Birnbaum, 1968) can be expressed as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{10}$$

where $P_i(\theta)$ is the probability of a correct response on the $i^{th}$ item for a subject with ability $\theta$. In this function $e$ is a constant ($e$ = 2.71828...) and $D$ is a scaling constant equal to 1.702. The parameter $c_i$ is the item pseudo-guessing parameter, $b_i$ is the item difficulty, and $a_i$ is proportional to the slope of the item characteristic curve at the inflection point and is, therefore, a discrimination parameter.

In the IRT model, once the item parameters are determined, the probability of a correct response is solely a function of examinee ability.

As examinee ability increases, the probability of a correct response also increases. Using Equation (10), the probability of a correct response can be used to generate an observable dichotomous right-wrong response by comparing it to a random number from a uniform distribution on the interval [0, 1]. A response was coded as correct when its associated probability was greater than the random number and incorrect when it was less. This procedure was used in the DATAGEN program by Hambleton and Rovinelli (1973) to generate item responses of a single group to an unbiased test.

For the study, one hundred unbiased data sets of a 34 item test were generated for both the reference and the focal groups. The $c$ parameters were all set at 0.20 which corresponds to the random guessing level for a five-option item. For the data simulation, $a_i$ and $b_i$ parameters were randomly selected from the normal (1, .3) and the normal (0, 1) distributions, respectively.

Shealy & Stout (1993) showed that the Mantel-Haenszel method yielded good adherence to the nominal significance levels even for differences in ability as large as one standard deviation, but Narayanan and Swaminathan (1994) argued that the difference in ability distribution had an effect on the MH method. This study used a one

standard deviation difference in ability between the reference and focal groups. The ability parameters for the reference group of 1,000 examinees were sampled from a normal distribution, $N(0, 1)$ while the ability parameters for the focal groups of 1,000, 500, and 200 examinees were sampled from a normal distribution, $N(-1, 1)$.

Simulated DIF

In the IRT framework, DIF exists when item characteristic curves (ICC) for two groups are different (Lord, 1980). Previous simulation studies (Donoghue, Holland, & Thayer; 1993; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; and Uttaro & Millsap; 1994) have used varying amounts of simulating DIF. This study, however, used only one level of DIF, an area of .4 between item characteristic curves (ICC) for the reference and focal groups. One hundred different data sets for the reference and the focal groups were generated to create one hundred replications for each type of data set.

In simulating uniform DIF data sets, the discrimination parameters for the two groups were held constant but the difficulty parameters were varied. Six types of uniform DIF were generated: (1) low $b$, low $a$; (2) low $b$ high $a$; (3) moderate $b$, low $a$; (4) moderate $b$, high $a$; (5) high $b$, low $a$; and (6) high $b$, low $a$.

Nonuniform DIF data sets included both symmetric and nonsymmetric DIF. In simulating symmetric nonuniform DIF, the difficulty parameters for the two groups were held constant but the discrimination parameters were varied. In simulating nonsymmetric nonuniform DIF, both the discrimination and the difficulty parameters for the two groups were varied. Six types of nonuniform DIF were generated with various combination of $b$ and $a$ as in the uniform DIF.

In simulating mixed DIF data sets, six types of DIF were generated: for uniform DIF (1) low $b$, moderate $a$; and (2) high $b$, moderate $a$; for symmetric nonuniform DIF (3) low $b$, low $a$; and (4) high $b$, low $a$; and for nonsymmetric nonuniform DIF (5) moderate $b$, low $a$; and (6) moderate $b$, high $a$. The conditions for generating DIF on each data set are shown in Table 3.

---

Insert Table 3 Here

---

Procedure

The chi-square statistics were computed in two steps as proposed by Holland and Thayer (1988). First, score groups were obtained using total scores based on all items, and then the chi-square statistics were

computed for all items. Those items for which the test statistic exceeded the critical value at $\alpha$=.05 or $\alpha$=.01 were identified and labeled as potentially displaying DIF. Next, total scores were reconstituted after eliminating items previously identified as DIF, and then the test statistics were calculated again.

## Results

This study was composed of two components. One is a false positive error study and the other is a power study. A one between two within factor repeated-measures design (Winer, 1962) was used to investigate the false positive rate and the power. In this design, the dependent variable was the false positive rate or the power, and the independent variables were methods (i.e., MH, UMH, and AMD), steps (i.e., with and without the purification procedure) and the sample sizesin the two groups (i.e., 1,000 vs 1,000, 1,000 vs 500, and 1,000 vs 200).

In each DIF condition, there was a total of 4,000 items. For the false positive study, the 3,400 unbiased items were used for the three types of data sets. For the power study, the 600 biased items were used for the three types of data sets. False positive rates and power were

computed separately for each type of DIF data set (uniform, nonuniform, or mixed) and each $\alpha$ level (.01 and .05).

False Positive Errors

Table 4 shows the ANOVA results of the one between two within factor repeated-measures design. Tables 5, 6, and 7 present the corresponding false positive rates and the number of items for each type of data set that were used in this analysis.

---

Insert Table 4, 5, 6, and 7 here

---

Table 4 shows that there were significant three-way interaction effects and two-way interactions although most main effects were significant. Usually, when there are significant higher order interactions, it is meaningless to discuss main effects. However, the descriptive statistics for all the effects may help to interpret the results.

Tables 5, 6 and 7 show that the two-step purification procedure reduced the false positive rates for the AMD method in most conditions but it did not for the MH and the UMH methods. Although the AMD method had the lower false positive error rates, it produced rates at $\alpha$ = .01 at least 1.5 times higher than the nominal $\alpha$ levels. The false positive

rates were worst in all conditions for the UMH. For instance, for the UMH in the 1,000 vs. 1,000 sample size comparison of nonuniform data sets at $\alpha = .01$, the false positive rates in the nonuniform DIF condition were 12 times as much as the nominal $\alpha$ levels for both before and after purification (See Table 6). In both steps (before and after purification), all methods tended to have false positive rates higher than the nominal $\alpha$ levels.

One noticeable result was that in both steps (before and after purification) the UMH method had high false positive error rates in all conditions. Additionally, the AMD method has the lowest false positive error rates in most conditions. The other noticeable result was that as sample sizes were decreased, the false positive error rates decreased for the MH method but it increased for the UMH method (See Tables 5, 6, and 7).

Power Study

Table 8 presents the ANOVA results of the one between two within factor repeated-measures design for the detection rates in each factor and at each level of $\alpha$. Tables 9, 10, and 11 show the corresponding mean detection rates and the number of items for uniform, nonuniform, and mixed data sets, respectively.

---

Insert Tables 8, 9, 10 and 11 here

---

Table 8 shows that there were significant main effects and interaction effects in all but two three-way interaction and two two-way interaction effects. All three-way and two-way interactions, excluding the mixed data set at $\alpha = .01$ and the nonuniform data set at $\alpha = .05$, were significant.

Tables 9, 10, and 11 show that the deterioration of power from step 1 (before purification) to step 2 (after purification) was larger for the MH method in the 1,000 vs. 500 sample size comparison of the uniform data sets for $\alpha = .05$ than for the others while the UMH and the AMD methods had larger differences between step 1 and step 2 in the 1,000 vs. 1,000 sample size comparison of the uniform DIF condition at $\alpha = .05$---that is, the deterioration rates were 60% for the MH method in the 1,000 vs. 500 sample size comparison of the uniform data sets for $\alpha = .05$; and it was 47% and 22% for the UMH and the AMD methods in the 1,000 vs. 1,000 sample size comparison of the uniform DIF condition at $\alpha = .05$, respectively.

The results show that the MH method had more power in step 1

(before purification) but the AMD method had more power in step 2 (after purification) of the 1,000 vs. 1,000 sample size comparison (See Tables 9, 10, and 11). As sample size decreased, the detection rate dropped dramatically for all the methods, especially on uniform DIF data sets.

## Discussion and Conclusion

As with any simulation study, conclusions are limited to the conditions in the simulation. Two particular features of this study are of note. First, ability distribution means were different for both the reference and focal groups. Second, the nonuniform DIF conditions were favorable to the MH method because all difficulty parameters for nonuniform DIF conditions were out of the mid-ranges of the combined ability distribution for both groups. Consequently, the ICC curves of the two groups crossed well above or below the mean of the combined distribution of two groups. Given these conditions,---unequal group means and absence of mid-range difficulty parameter for both groups---the MH canceling effect for nonuniform DIF is minimal.

In simulating DIF, there are two versions of the null hypothesis which vary as a function of the context formed by the other items in the

test when, as is usually the case, the test is the matching variable. First, the null hypothesis is true for the studied item and true for all items in the test constituting the matching variable. Second, the null hypothesis is true for the studied item, but false for one or more of the other items in the test comprising the matching variable. Our findings regarding false positives apply only to the second of these two situations.

The rationale for the purification approach is that items with DIF will degrade ability estimation, which in turn may adversely affect the detection of DIF. When ability distributions for both groups are the same, only DIF can "contaminate" the items leading to identification in the first step of the purification procedure. Therefore, the removal of items with DIF may reduce the false positives or increase power. Previous simulation studies (Kwak, 1994; Kwak, Davenport, & Davison, 1997; Miller, & Oshima, 1992) supported this argument. However, when ability distributions differ, DIF may be compounded with impact. Non-DIF items may be removed in the first step of the purification procedure because they contain impact. Therefore, the removal of items identified in the first step may result in lower power for the second step of the purification procedure.

Three major conclusions arise from this study. First, when the ability distributions for the two groups are different, the two-step

purification procedure increases the false positive error rates for the MH and the UMH methods and reduces the power for all the three methods. Second, the Mantel-Haenszel method was more powerful than either the absolute mean deviation statistic or the unsigned Mantel-Haenszel method for detecting uniform DIF, and nonuniform DIF when the interaction occurs outside of the middle range of the combined ability distribution for the two groups. Third, as sample size decreases, the detection rate of DIF also decreases.

In the past, uniform DIF has been of the greatest concern to researchers. However, nonuniform DIF has emerged in empirical data (Bennett, Rock, & Kaplan, 1987; Ellis, 1989; Hambleton & Rogers, 1989; Mellenbergh, 1983) and Millsap (1995) shows how nonuniform DIF can readily emerge in practical applications. Nonuniform DIF cannot be ignored. In our data, both the absolute mean deviation statistic and the MH method yielded reasonably high detection rates of nonuniform DIF.

Given our findings, those using chi-square based methods may want to combine the Mantel-Haenszel with the absolute mean deviation statistic. The latter seems to provide smaller false positive error rates and may even be a similar or more powerful test when nonuniform DIF exists.

## References

Agresti, A. (1990). *Categorical data analysis*. NY: John Wiley & Sons.

Agresti, A. (1996). *An introduction to categorical data analysis*. NY: John Wiley & Sons.

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

Birnbaum, A. (1968). 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* (pp. 397-479). Reading, MA: Addison-Wesley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Clauser, B. E., Nungster, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*, 202-214.

Donoghue, J., Holland, P. W., & Thayer, D. T. (1993). A Monte

Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hiilsdale, NJ: Erlbaum.

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

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.

Hambleton, R. K., & Rovinelli, R. (1973). A FORTRAN program for generating examinee response data from logistic test models. *Behavioral Science, 18,* 74.

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

Kwak, N. (1994). *A simulation study of methods for detecting uniform and nonuniform item bias*. Unpublished Master Thesis, University of Minnesota.

Kwak, N., Davenport, Jr. E. C., & Davison, M. L. (1997). A comparison of the absolute mean deviation statistic and two selected chi-square statistics for detecting uniform and nonuniform DIF. Manuscript submitted for publication.

Li, H-, H, & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

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

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

Mellenbergh, G. J. (1983). Conditional item bias methods. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cultural factors* (pp. 293-302). NY: Plenum Press.

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

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*, 381-388.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*, 577-605.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Shealy, R., & Stout, W. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and

detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

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

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detecting of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

Winer, B. J. (1971). *Statistical principles in experimental design*. NY: McGraw-Hill.

Table 1

Data for the *j*th Matched Set of Members of High Proportion Correct (H) and Low Proportion Correct (L) Groups on the Studied Item

| Group | Score on Studied Item 1 | Score on Studied Item 0 | Total |
|---|---|---|---|
| $H$ | $H_{1i}$ | $H_{0i}$ | $N_{Hi}$ |
| $L$ | $L_{1i}$ | $L_{0i}$ | $N_{Li}$ |
| Total | $N_{1i}$ | $N_{0i}$ | $N_i$ |

Table 2

Data for the *j*th Matched Set of Members of Reference (R) and Focal (F) Groups on the Studied Item

| Group | Score on Studied Item 1 | Score on Studied Item 0 | Total |
|---|---|---|---|
| $R$ | $A_i$ | $B_i$ | $N_{Ri}$ |
| $F$ | $C_i$ | $D_i$ | $N_{Fi}$ |
| Total | $N_{1i}$ | $N_{0i}$ | $N_i$ |

Table 3

Item Parameters Used to Generate Items with DIF (*c* Values for Both Reference and Focal Groups Were Fixed at .2.)

| Data Set with DIF Condition of Item Parametr | Reference | | Focal | |
|---|---|---|---|---|
| | $b_1$ | $a_1$ | $b_2$ | $a_2$ |
| Uniform | | | | |
| Low *b*, Low *a* | -1.70 | 0.70 | -1.12 | 0.70 |
| High *a* | -1.70 | 1.30 | -1.18 | 1.30 |
| Moderate *b*, Low *a* | -0.30 | 0.70 | 0.23 | 0.70 |
| High *a* | -0.30 | 1.30 | 0.20 | 1.30 |
| High *b*, Low *a* | 1.30 | 0.70 | 1.90 | 0.70 |
| High *a* | 1.30 | 1.30 | 1.82 | 1.30 |
| Nonuniform | | | | |
| Low *b*, Low *a* | -1.40 | 0.32 | -1.30 | 0.56 |
| High *a* | -1.76 | 1.20 | -1.24 | 1.85 |
| Moderate *b*, Low *a* | 0 | 0.42 | 0 | 0.69 |
| High *a* | 0 | 1.65 | 0 | 0.80 |
| High *b*, Low *a* | 1.30 | 0.35 | 1.40 | 0.59 |
| High *a* | 1.25 | 1.20 | 1.75 | 1.56 |
| Mixed | | | | |
| Low *b*, Moderate *a* | -1.74 | 1.00 | -1.20 | 1.00 |
| High *b* | 1.20 | 1.00 | 1.74 | 1.00 |
| Low *b*, Low *a* | -1.50 | 0.41 | -1.50 | 0.73 |
| High *b* | 1.50 | 0.41 | 1.50 | 0.73 |
| Moderate *b*, Low *a* | -0.05 | 0.32 | 0.05 | 0.53 |
| High *a* | -0.20 | 1.00 | 0.20 | 1.83 |

Table 4

F Ratio and p Value of the One Between Two Within Factor Repeated-Measure Design for the False Positive Rate

| Data Set with DIF | Factor | $\alpha = .01$ F | $\alpha = .01$ p | $\alpha = .05$ F | $\alpha = .05$ p |
|---|---|---|---|---|---|
| Uniform | S | 3.27 | .038 | 3.38 | .034 |
| | P | 40.02 | .000 | 179.73 | .000 |
| | M | 445.65 | .000 | 837.00 | .000 |
| | S × P | 3.44 | .032 | 2.69 | .068 |
| | S × M | 28.01 | .000 | 49.17 | .000 |
| | P × M | 64.25 | .000 | 131.54 | .000 |
| | S × P × M | 5.94 | .000 | 19.68 | .000 |
| Nonuniform | S | 14.21 | .000 | 12.97 | .000 |
| | P | 4.00 | .045 | 77.21 | .000 |
| | M | 702.83 | .000 | 1361.26 | .000 |
| | S × P | 10.84 | .000 | 12.70 | .000 |
| | S × M | 16.21 | .000 | 49.64 | .000 |
| | P × M | 44.08 | .000 | 126.75 | .000 |
| | S × P × M | 7.07 | .000 | 11.25 | .000 |
| Mixed | S | 1.06 | .347 | 3.96 | .019 |
| | P | 37.65 | .000 | 216.44 | .000 |
| | M | 570.28 | .000 | 1099.77 | .000 |
| | S × P | 11.65 | .000 | 13.79 | .000 |
| | S × M | 25.36 | .000 | 65.13 | .000 |
| | P × M | 73.71 | .000 | 153.22 | .000 |
| | S × P × M | 10.14 | .000 | 31.47 | .000 |

S: sample size
P: two-step purification procedure
M: method

Table 5

The False Positive Rate of Uniform DIF Data Set for Factors at Each Level of α

| Sample Size | Purification and Method | α = .01 p | α = .01 N | α = .05 p | α = .05 N |
|---|---|---|---|---|---|
| 1000 vs 1000 | Step 1 | | | | |
| | MH | .030 | 3,400 | .087 | 3,400 |
| | UMH | .058 | 3,400 | .146 | 3,400 |
| | AMD | .020 | 3,400 | .063 | 3,400 |
| | Step 2 | | | | |
| | MH | .031 | 3,400 | .099 | 3,400 |
| | UMH | .067 | 3,400 | .174 | 3,400 |
| | AMD | .017 | 3,400 | .070 | 3,400 |
| 1000 vs 500 | Step 1 | | | | |
| | MH | .021 | 3,400 | .073 | 3,400 |
| | UMH | .059 | 3,400 | .146 | 3,400 |
| | AMD | .019 | 3,400 | .069 | 3,400 |
| | Step 2 | | | | |
| | MH | .021 | 3,400 | .077 | 3,400 |
| | UMH | .071 | 3,400 | .189 | 3,400 |
| | AMD | .018 | 3,400 | .072 | 3,400 |
| 1000 vs 200 | Step 1 | | | | |
| | MH | .012 | 3,400 | .058 | 3,400 |
| | UMH | .083 | 3,400 | .190 | 3,400 |
| | AMD | .021 | 3,400 | .069 | 3,400 |
| | Step 2 | | | | |
| | MH | .014 | 3,400 | .061 | 3,400 |
| | UMH | .108 | 3,400 | .265 | 3,400 |
| | AMD | .015 | 3,400 | .059 | 3,400 |

Table 6

The False Positive Rate of Nonuniform DIF Data Set for Factors at Each Level of α

| Sample Size | Purification and Method | α = .01 | | α = .05 | |
|---|---|---|---|---|---|
| | | p | N | p | N |
| 1000 vs 1000 | Step 1 | | | | |
| | MH | .071 | 3,400 | .178 | 3,400 |
| | UMH | .122 | 3,400 | .252 | 3,400 |
| | AMD | .028 | 3,400 | .094 | 3,400 |
| | Step 2 | | | | |
| | MH | .071 | 3,400 | .187 | 3,400 |
| | UMH | .126 | 3,400 | .281 | 3,400 |
| | AMD | .025 | 3,400 | .075 | 3,400 |
| 1000 vs 500 | Step 1 | | | | |
| | MH | .046 | 3,400 | .130 | 3,400 |
| | UMH | .098 | 3,400 | .234 | 3,400 |
| | AMD | .027 | 3,400 | .086 | 3,400 |
| | Step 2 | | | | |
| | MH | .044 | 3,400 | .138 | 3,400 |
| | UMH | .105 | 3,400 | .262 | 3,400 |
| | AMD | .016 | 3,400 | .074 | 3,400 |
| 1000 vs 200 | Step 1 | | | | |
| | MH | .026 | 3,400 | .085 | 3,400 |
| | UMH | .106 | 3,400 | .253 | 3,400 |
| | AMD | .019 | 3,400 | .072 | 3,400 |
| | Step 2 | | | | |
| | MH | .025 | 3,400 | .085 | 3,400 |
| | UMH | .129 | 3,400 | .320 | 3,400 |
| | AMD | .016 | 3,400 | .071 | 3,400 |

Table 7

The False Positive Tate of Mixed DIF Data Set for Factors at Each Level of α

| Sample Size | Purification and Method | α = .01 | | α = .05 | |
|---|---|---|---|---|---|
| | | p | N | p | N |
| 1000 vs 1000 | Step 1 | | | | |
| | MH | .044 | 3,400 | .129 | 3,400 |
| | UMH | .080 | 3,400 | .190 | 3,400 |
| | AMD | .023 | 3,400 | .071 | 3,400 |
| | Step 2 | | | | |
| | MH | .046 | 3,400 | .142 | 3,400 |
| | UMH | .084 | 3,400 | .214 | 3,400 |
| | AMD | .016 | 3,400 | .075 | 3,400 |
| 1000 vs 500 | Step 1 | | | | |
| | MH | .032 | 3,400 | .095 | 3,400 |
| | UMH | .074 | 3,400 | .181 | 3,400 |
| | AMD | .019 | 3,400 | .069 | 3,400 |
| | Step 2 | | | | |
| | MH | .033 | 3,400 | .102 | 3,400 |
| | UMH | .095 | 3,400 | .227 | 3,400 |
| | AMD | .018 | 3,400 | .069 | 3,400 |
| 1000 vs 200 | Step 1 | | | | |
| | MH | .020 | 3,400 | .069 | 3,400 |
| | UMH | .097 | 3,400 | .224 | 3,400 |
| | AMD | .021 | 3,400 | .075 | 3,400 |
| | Step 2 | | | | |
| | MH | .019 | 3,400 | .074 | 3,400 |
| | UMH | .126 | 3,400 | .318 | 3,400 |
| | AMD | .017 | 3,400 | .069 | 3,400 |

Table 8

F Ratio and p Value of the One Between Two Within Factor Repeated-Measure Design for Power

| Data Set with DIF | Factor | α = .01 F | α = .01 p | α = .05 F | α = .05 p |
|---|---|---|---|---|---|
| Uniform | S | 160.98 | .000 | 224.01 | .000 |
| | P | 275.93 | .000 | 494.12 | .000 |
| | M | 438.87 | .000 | 629.45 | .000 |
| | S × P | 39.96 | .000 | 40.04 | .000 |
| | S × M | 31.98 | .000 | 23.60 | .000 |
| | P × M | 41.50 | .000 | 91.38 | .000 |
| | S × P × M | 4.06 | .003 | 7.56 | .000 |
| Nonuniform | S | 98.72 | .000 | 105.62 | .000 |
| | P | 2.74 | .098 | 5.79 | .016 |
| | M | 556.86 | .000 | 590.64 | .000 |
| | S × P | 3.65 | .026 | 1.30 | .274 |
| | S × M | 31.80 | .000 | 43.09 | .000 |
| | P × M | .01 | .989 | 8.79 | .000 |
| | S × P × M | 1.96 | .098 | .81 | .521 |
| Mixed | S | 172.86 | .000 | 176.32 | .000 |
| | P | 59.12 | .000 | 140.86 | .000 |
| | M | 589.85 | .000 | 568.50 | .000 |
| | S × P | 1.28 | .278 | 15.42 | .000 |
| | S × M | 35.70 | .000 | 53.36 | .000 |
| | P × M | 11.12 | .000 | 21.22 | .000 |
| | S × P × M | 1.18 | .318 | 3.32 | .010 |

S: sample size
P: two-step purification procedure
M: method

Table 9

The Power of Uniform DIF Data Set for Factors at Each Level of α

| Sample Size | Purification and Method | α = .01 p | α = .01 N | α = .05 p | α = .05 N |
|---|---|---|---|---|---|
| 1000 vs 1000 | Step 1 | | | | |
| | MH | .573 | 600 | .810 | 600 |
| | UMH | .372 | 600 | .587 | 600 |
| | AMD | .410 | 600 | .647 | 600 |
| | Step 2 | | | | |
| | MH | .343 | 600 | .468 | 600 |
| | UMH | .205 | 600 | .313 | 600 |
| | AMD | .348 | 600 | .492 | 600 |
| 1000 vs 500 | Step 1 | | | | |
| | MH | .373 | 600 | .620 | 600 |
| | UMH | .165 | 600 | .318 | 600 |
| | AMD | .312 | 600 | .503 | 600 |
| | Step 2 | | | | |
| | MH | .237 | 600 | .350 | 600 |
| | UMH | .075 | 600 | .152 | 600 |
| | AMD | .275 | 600 | .363 | 600 |
| 1000 vs 200 | Step 1 | | | | |
| | MH | .172 | 600 | .357 | 600 |
| | UMH | .020 | 600 | .075 | 600 |
| | AMD | .037 | 600 | .207 | 600 |
| | Step 2 | | | | |
| | MH | .082 | 600 | .190 | 600 |
| | UMH | .015 | 600 | .045 | 600 |
| | AMD | .023 | 600 | .182 | 600 |

Table 10

The Power of Nonuniform DIF Data Set for Factors at Each Level of α

| Sample Size | Purification and Method | α = .01 p | α = .01 N | α = .05 p | α = .05 N |
|---|---|---|---|---|---|
| 1000 vs 1000 | Step 1 | | | | |
| | MH | .695 | 600 | .782 | 600 |
| | UMH | .543 | 600 | .627 | 600 |
| | AMD | .662 | 600 | .778 | 600 |
| | Step 2 | | | | |
| | MH | .700 | 600 | .760 | 600 |
| | UMH | .552 | 600 | .640 | 600 |
| | AMD | .667 | 600 | .782 | 600 |
| 1000 vs 500 | Step 1 | | | | |
| | MH | .640 | 600 | .725 | 600 |
| | UMH | .417 | 600 | .495 | 600 |
| | AMD | .598 | 600 | .673 | 600 |
| | Step 2 | | | | |
| | MH | .640 | 600 | .677 | 600 |
| | UMH | .378 | 600 | .490 | 600 |
| | AMD | .565 | 600 | .667 | 600 |
| 1000 vs 200 | Step 1 | | | | |
| | MH | .505 | 600 | .620 | 600 |
| | UMH | .157 | 600 | .203 | 600 |
| | AMD | .343 | 600 | .547 | 600 |
| | Step 2 | | | | |
| | MH | .483 | 600 | .558 | 600 |
| | UMH | .162 | 600 | .232 | 600 |
| | AMD | .348 | 600 | .535 | 600 |

Table 11

The Power of Mixed DIF Data Set for Factors at Each Level of α

| Sample Size | Purification and Method | α = .01 p | α = .01 N | α = .05 p | α = .05 N |
|---|---|---|---|---|---|
| 1000 vs 1000 | Step 1 | | | | |
| | MH | .682 | 600 | .817 | 600 |
| | UMH | .517 | 600 | .673 | 600 |
| | AMD | .583 | 600 | .778 | 600 |
| | Step 2 | | | | |
| | MH | .597 | 600 | .652 | 600 |
| | UMH | .475 | 600 | .517 | 600 |
| | AMD | .570 | 600 | .671 | 600 |
| 1000 vs 500 | Step 1 | | | | |
| | MH | .640 | 600 | .678 | 600 |
| | UMH | .417 | 600 | .393 | 600 |
| | AMD | .598 | 600 | .653 | 600 |
| | Step 2 | | | | |
| | MH | .640 | 600 | .580 | 600 |
| | UMH | .378 | 600 | .298 | 600 |
| | AMD | .565 | 600 | .578 | 600 |
| 1000 vs 200 | Step 1 | | | | |
| | MH | .508 | 600 | .532 | 600 |
| | UMH | .157 | 600 | .060 | 600 |
| | AMD | .343 | 600 | .463 | 600 |
| | Step 2 | | | | |
| | MH | .483 | 600 | .433 | 600 |
| | UMH | .162 | 600 | .058 | 600 |
| | AMD | .348 | 600 | .460 | 600 |

AREA 1997 TM026903

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

# REPRODUCTION RELEASE
(Specific Document)

## I. DOCUMENT IDENTIFICATION:

| | |
|---|---|
| Title: An Unsigned Mantel-Haenszel Statistic for Detecting Uniform and Nonuniform DIF | |
| Author(s): Nohoon Kwak; Mark L. Davison; Ernest C. Davenport Jr. | |
| Corporate Source: University of Minnesota | Publication Date: 3/27/97 |

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

[X] Sample sticker to be affixed to document

**Check here**
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

**Level 1**

Sample sticker to be affixed to document [ ]

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

**Level 2**

**or here**
Permitting reproduction in other than paper copy.

## Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

| | |
|---|---|
| Signature: Nohoon Kwak | Position: |
| Printed Name: Nohoon Kwak | Organization: University of Minnesota |
| Address: Dept. of Educational Psychology<br>University of Minnesota<br>178 Pillsbury Drive S.E.<br>Minneapolis, MN 55455 | Telephone Number: (612) 626-7881 |
| | Date: 4/14/97 |

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
*Department of Education, O'Boyle Hall*
*Washington, DC 20064*
*202 319-5120*

February 21, 1997

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an AERA chair or discussant, please save this form for future use.