p- ISSN: 2708-2113 e-ISSN: 2708-3608 L-ISSN: 2708-2113 URL: http://dx.doi.org/10.31703/gesr.2021(VI-III).08

Vol. VI, No. III (Summer 2021) Pages: 71– 78 DOI: 10.31703/gesr.2021(VI-III).08

Citation: Khalid, M. N., Shafiq, F., & Ahmed, S. (2021). Detection of Differential Item Functioning Using Mantel-Haenszel, Standardization Proportion and BILOG-MG Procedures. *Global Educational Studies Review, VI*(III), 71-78. https://doi.org/10.31703/gesr.2021(VI-III).08

Detection of Differential Item Functioning Using Mantel-Haenszel, Standardization Proportion and BILOG-MG Procedures

Muhammad Naveed Khalid [*] | Farah Shafiq [†] | Shehzad Ahmed [‡]

Abstract: *Differential item functioning (DIF) is a procedure to identify whether an item favours a particular group of respondents once they are matched on respective ability levels. There are numerous procedures reported in the literature to detect DIF, but the Mantel-Haenszel (MH), Standardized Proportion Difference (SPD), and BILOG-MG are frequently used to ensure the fairness of assessments. The aim of the present study was to compare procedural characteristics using empirical data. We found Mantel-Haenszel and standardized proportion difference provide comparable results while BILOG-MG has flagged a large number of items, but the magnitude of DIF was trivial from a test development perspective. The results also showed Mantel-Haenszel and standardized proportion difference index provide the effect size measure of DIF, which facilitates for further necessary actions, especially for item writers and practitioners.*

Key Words: Differential Item Functioning, Effect Size, Classification, MH, SPD, BILOG-MG

Introduction

Differential item functioning (DIF) is a statistical procedure to examine the fairness of assessment across various groups of respondents. It allows us to analyze the item performance in the groups of interest once they are matched on the overall ability (Holland & Wainer, 1993). An item shows DIF if respondents who belong to different groups such as gender, demographics, and age group and have the same ability level but the probability of choosing a correct option is different (Millsap & Everson, 1993). In DIF analysis, we compare the item difficulty in two groups of respondents. These groups are named as a reference group and the focal group, and they may appear as uniform or non-uniform. Interaction of group membership and ability of respondents does not occur in uniform DIF, while in the case of non-uniform DIF there is an interaction (Mellenbergh, 1982).

There are numerous procedures for the detection of DIF in both objective and constructed response items (Hidalgo & Gómez-Benito, 2010). But the Mantel-Haenszel (MH) standardized proportion difference (SPD) statistic are observed as a reference technique because their computation details are quite straightforward and can be applied to small samples, and their characteristics are well studied in the DIF literature (Guilera, Gómez-Benito & Hidalgo, 2009). A related technique that also has been applied and studied widely is the critical ratio test which is implemented in BILOG-MG.

The objective of the current investigation is to study the characteristics of the above-mentioned procedures. We conducted a comparative analysis of these procedures using an empirical dataset. The outline of the study is as follows. The first section presents a brief description of the DIF statistics. The next section will sketch the idea of scale purification and effect size. Then, the description of the data and results of the study are tabulated. Finally, some conclusions are drawn for applied settings.

[*] Resource Person, Allama Iqbal Open University, Lahore, Punjab, Pakistan. Email: naveedscholar@gmail.com
[†] Assistant Professor, Department of Education, University of Education, Lahore, Punjab, Pakistan.
[‡] Assistant Professor, Faculty of Education, University of Okara, Punjab, Pakistan.

Differential Item Functioning Statistics

Mantel-Haenszel Statistics

The MH statistical examines the probability of giving a correct response in two groups (reference and focal) after they are matched on the same ability level (Holland and Thayer, 1988). For each item contingency analysis performed at each score level to examine whether DIF is present, an example of such analysis is shown in Table 1. Let's assume that there are $N_{..j}$ examinees at the *jth* level. $N_{R.j}$ and $N_{F.j}$ denote the number of respondents belonging to reference and focal group. A_j and B_j show how many students gave a correct and incorrect answers. Likewise, C_j out of the $N_{F.j}$ answered correctly, whereas D_j did not. A "." denotes summation over a particular index.

Table 1. Correct Number Score on *ith* Item in *j* Score

	Item Score		
Group	**1**	**0**	**Total**
Reference	A_j	B_j	$N_{R.j}$
Focal	C_j	D_j	$N_{F.j}$
Total	$N_{1.j}$	$N_{0.j}$	$N_{..j}$

The odds ratio to compute DIF between groups of interest-based on say, at score level *j* is given by:

$$\alpha = \left(p_{Rj}/1-p_{Rj}\right)/\left(p_{Fj}/1-p_{Fj}\right) \quad (1)$$

in which p_{Rj} and p_{Fj} are the probabilities to choose a particular option for the respective groups. These probabilities at score level *j* are computed in the following manner:

$$p_{Fj} = \frac{C_j}{N_{F.j}} \text{ And } p_{Rj} = \frac{A_j}{N_{R.j}}.$$

The MH statistic for an item exhibiting DIF computed as:

$$MH = \frac{\left[\left|\sum_{j=1}^{K} A_j - \sum_{j=1}^{K} E(A_j)\right| - 0.5\right]^2}{\sum_{j=1}^{K} Var(A_j)} \quad (2)$$

where as $E(A_j) = \frac{N_{R.j}N_{1.j}}{N_{..j}}$ (3)

$$Var(A_j) = \frac{N_{R.j}N_{F.j}N_{1.j}N_{0.j}}{(N_{..j})^2(N_{..j}-1)}$$

The distribution of the MH statistic is χ^2 distribution with one degree of freedom and effect size based on the common odds ratio α is expressed as

$$\alpha_{MH} = \frac{\sum_{j=1}^{K} A_j D_j / N_{..j}}{\sum_{j=1}^{K} B_j C_j / N_{..j}} \quad (4)$$

A delta metric scale of α is suggested by Holland and Thayer (1988), which is given in equation 5. The item gives an advantage to the reference group whose delta value is negative, while a positive value shows DIF in the opposite direction. Similarly, Zwick and Ercikan (1989) proposed guidelines to classify the items based on the magnitude of DIF as Type A, Type B, and Type C items.

$$\Delta\alpha_{MH} = -2.35 \ln(\alpha_{MH}) \quad (5)$$

- Type A items - negligible DIF: items with $|\Delta\alpha_{MH}| < 1$ or *MH* test is not statistically significant and considered to function properly.
- Type B items - moderate DIF: items with $1 \leq |\Delta\alpha_{MH}| \leq 1.5$, and *MH* test is statistically significant. They could be used who have the lowest $\Delta\alpha_{MH}$ values and do not have alternative items.
- Type C items - large DIF: items with $|\Delta\alpha_{MH}| > 1.5$, and *MH* test is statistically significant. A critical review of these items is necessary and will be only selected in exceptional circumstances.

Standardization Procedure

The standardization procedure is based on the comparison of the item score in different groups once the level of the ability has been matched.

Dorans and Holland (1993) formulated the SPD index (Standardized Proportions Difference) to identify DIF by using conditional proportions. The SPD index is denoted by

$$SPD = \frac{\sum_{j=1}^{K} w_j \left(p_{Fj} - p_{Rj}\right)}{\sum_{j=1}^{k} w_j} \qquad (6)$$

In which $p_{Fj} = \frac{C_j}{N_{Fj}}$ $p_{Rj} = \frac{A_j}{N_{Rj}}$ and are the success proportions in the item for the respondents in the *j* stratus in focal and reference group respectively, and w_j is a weight factor (standardization parameter) of the difference in this score level.

The weight factor is is one of the essential elements in the procedure because it distinguishes the DIF computation from the computation of impact. Some values that w_j can take are: a) N_j: Total number of respondents in the stratus j, b) N_{Rj}: Number of respondents in the reference group in the stratus j, c) N_{Fj}: Number of respondents in the focal group in the stratus *j*, or d) the relative frequency in the reference group in the stratus j. One of the most used is the number of respondents in the focal group in specific stratus (N_{Fj}) because it gives the greatest weight to (Dorans & Kulick, 1986).

The SPD index ranges from -1 to +1, and the positive value of the index favours the focal group. Dorans and Holland (1993) proposed values between -0.05 and 0.05 are considered as negligible; 0.05 and 0.1 in absolute value are doubtful, and values beyond the previous criteria point out a careful revision of the items.

DIF Detection in BILOG-MG

BILOG-MG is a program to examine DIF whether item difficulty is the same in the studied groups (Zimowski, Muraki, Mislevy, & Bock, 1996). Detection of DIF depends on the difference in the difficulty (b) parameter and a model comparison statistics -2 log-likelihood ratio (-2lnLR) test as expressed in equation 7. The difference shows how well models fit the data.

$$\chi^2(M) \approx -2\ln(LR) = G(2) - G(1) \qquad (7)$$

where

$M = df$, $G(2) = -2\ln L(\text{Model 2})$, $G(1) = -2\ln L(\text{Model 1})$

The program computes difficulty parameters across groups and receptive standard errors (*s.e.*) to determine whether the difficulty parameter is the same.

$$s.e._{G2-G1} = \sqrt{s.e.^2{}_{G2} + s.e.^2{}_{G1}} \qquad (8)$$

A statistical test called the critical ratio test can be computed based on difficulty parameters by dividing the *s.e.* for each item. Muraki and Engelhard's (1989) proposed larger than two standard deviations criteria to judge whether item exhibited DIF.

$$\text{Critical ratio test} = \frac{|\hat{b}_{Fi} - \hat{b}_{Ri}|}{\sqrt{s.e.^2{}_{G2} + s.e.^2{}_{G1}}} \qquad (9)$$

Where R represents the reference group, and F represents the focal group.

Levels of Matching Variable

The DIF research shows there are two matching strategies named as thin matching and thick matching. The former uses the total score while in a thick matching strategy based on dividing the total scores into equal intervals to form a matching variable. Donoghue and Allen (1993) described the main strategies used to form matching variables. They found for short tests (5 or 10 items), thin matching performed poorly in detecting DIF while yielding the best results for longer and intermediate test lengths.

Scale Purification

The DIF statistical procedures use the number correct score as their matching criterion. The caveat in forming matching variables on test scores is it may lead to inaccurate ability estimates that may flag non-DIF items as DIF (Kim & Cohen, 1992). To avoid this issue, various purification procedures have been proposed, and their details can be found in (Holland & Thayer, 1988; Candell & Drasgow, 1988; Hidalgo and Gómez-Benito, 2003; Clauser, Mazor and Hambleton, 1993).

Effect Size

The effect size of DIF is important to prevent flagging items as DIF where the magnitude of DIF is trivial but statistically significant results. It helps not to flag unimportant differences in large

samples and underestimation of DIF in small samples. Effect sizes also help to review test items from a test development perspective whether to retain or discard items.

Data

The example pertains to a standardized scale to assess the in-depth English language ability of respondents. It consists of five papers: Reading, Writing, Use of English, Listening, and Speaking. The listening section is used for the present investigation, and it comprises of 32 questions which include multiple-choice, sentence completion, and multiple matching in four sections. Each question was marked dichotomously as 0 and 1. The scale was administered to the students seeking admission to universities across the country. The sample consisted of 4865 respondents, and the respondents were divided into two groups based on their age.

Results

Descriptive Statistics

Descriptive measures of scores in each age group are presented in the below table. This investigation explored the item performance differences between the age groups. Examinees that have age 18-22 form the reference group while focal group comprised who have age 17 & under. The appeal and value of studied procedures cab be applied to any background variables of interest for investigating DIF, such as gender, race/ethnicity, and location or academic major.

Table 2. Descriptive Statistics

Age Group	Sample Size	Mean	SD	Reliability
18-22	2741	20.02	5.25	0.78
17 & Under	2124	18.00	5.33	0.78
Overall	4865	19.16	5.37	0.79

As can be seen from Table 2, the majority of examinees were between 18 and 22 years of age. The 18-22 age group candidates scored higher than candidates who belong to the 17 & under age group. Similarly, there is slightly more variation in the scores of focal group. The reliability of test scores in each age group and over all is similar.

The relative frequency of test scores in each age group is also shown in Figure 1. *X*-axis denotes the number correct score, and *Y*-axis shows the number of cases in each age group at each score point. Distributions of two groups are approximately normal; besides, the reference group is relatively more negatively skewed.

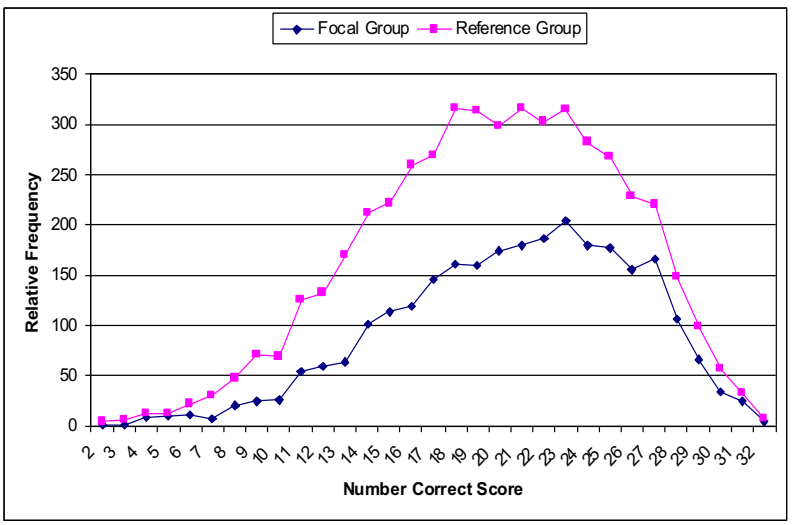

Figure 1: Graph of Total Score Distributions

DIF Detection using Mantel-Haenszel

Mantel-Haenszel chi-square statistics and delta values for each item were computed to flag the misfit ones. For the studied data set, all items functioned equally among the age groups. Further more, the magnitude of the effect size of items shows negligible DIF (A – Category). According to Holland and Thayer (1988), items should be flagged for further scrutiny that belongs to category C and B. The MH procedure was run in succession using thin and thick (Min. Freq of 5) matching strategies, but no difference is found.

MH-Delta index for each item is shown in figure 2. The item will be easier for the respondents of the references group if MH values are negative, while a positive value shows item favours in the opposite direction. It can be easily seen that the delta index associated with the item is within the A-category threshold. It can be concluded that candidates in each age group performed similarly on a test of listening. Magnitude of DIF and classification determined as per guidance given by Zwick and Ercikan (1989).

Table 3. Classification of DIF Items using MH Delta Index

Matching Level	Category - C	Category - B	Category – A
Thin	-	-	32
Thick (Min. Freq)	-	-	32

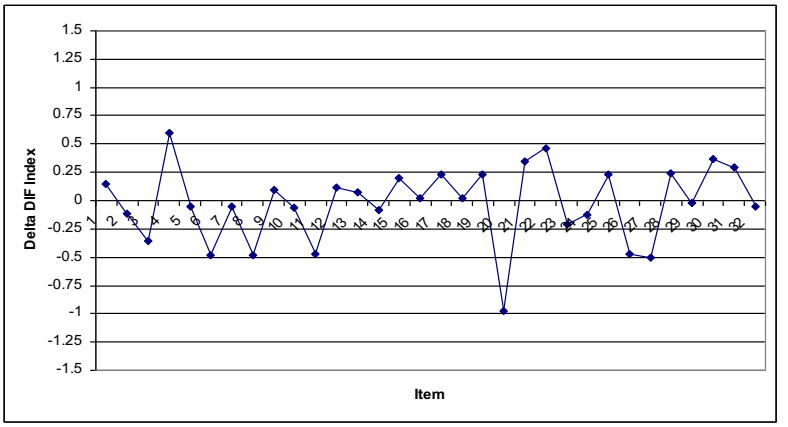

Figure 2: Presentation of the Complete set of MH-Delta Indices

DIF Detection using Standardized Proportion Difference

The standardized Proportion Difference index for each item was computed using a number of respondents at specific score level (N_{Fj}) who belong to the focal group as a weighting factor because it gives the greatest weight to differences (Dorans & Kulick, 1986). Overall, items functioned similarly across examinees irrespective of their age groups except for two items (4, 20). The DIF magnitude of these items was 0.06 and -0.07. The severity of DIF magnitude did not demand intensive revision or consideration. Table 4 shows the classification of items according to the magnitude of DIF. Furthermore, a high similarity is observed to flag and classify DIF items between Mantel-Haenszel and standardized proportion difference procedures.

Table 4. Classification of DIF Items using SPD Index

Between -0.05 and 0.05	Between 0.05 and 0.1	Between 0.1 or over
30	2	-

SPD index for each item is shown in figure 3. The item favours the focal group when the value is positive, and a negative value indicates DIF in the opposite direction. Values between -0.05 and 0.05 are considered as negligible; values between 0.05 and 0.1 in absolute value are doubtful, and values beyond the previous criterions point out a careful revision of the items. Figure 3 revealed a similar conclusion, as shown in Table 6. From figure 3, it can be easily seen that only two items, 4 and 20, have exceeded the threshold of ± 0.05 to be considered as doubtful items.

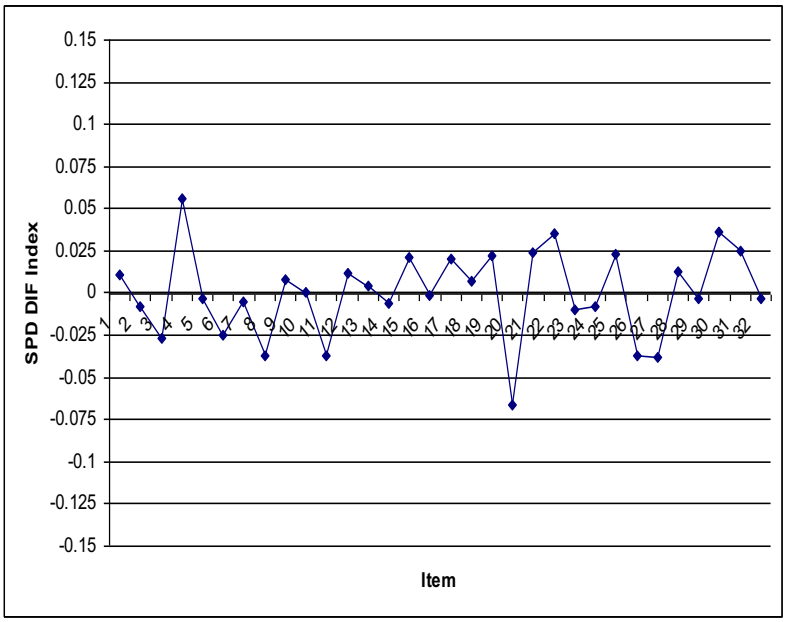

Figure 3: Presentation of the Complete set of SPD Indices

DIF Detection using BILOG-MG

BILOG-MG software produces tables with threshold differences between two groups under consideration and standard errors for each item. An item exhibits DIF if difference in difficulty level is beyond two standard deviations among the studied groups. For the studied data set, the magnitude of critical ratio, beyond ±2, shows 16 out of 32 items functioned differentially among the age groups. Table 5 shows 16 items demonstrate differential performance, 8 items favor reference group candidates, and 8 items favor examinees that belong to focal group.

Table 5. DIF Items using BILOG-MG

3, 4, 6, 8, 9, 11, 18, 19, 20, 21, 22, 24, 25, 26, 27, 30

Values of the critical ratio associated with each item is shown in figure 4. Interpretation of DIF magnitude (positive and negative) is similar to that we have discussed in the above sections. There is a large discrepancy in flagging DIF items between BILOG-MG and MH & SPD procedures. A large number of items are flagged as DIF (practically trivial but statistically significant) items. This inflation can be explained as follows.

Sample size played a crucial role in the computation of the standard errors. The magnitude of error is quite small for the large sample, which inflated the values of critical ratio test. In the literature, it is reported extensively that in the presence of a large sample size, even the smallest difference could be significant. Due to this behaviour of statistics, a large number of items will show DIF, while few items will exhibit DIF in the presence of a small sample size.

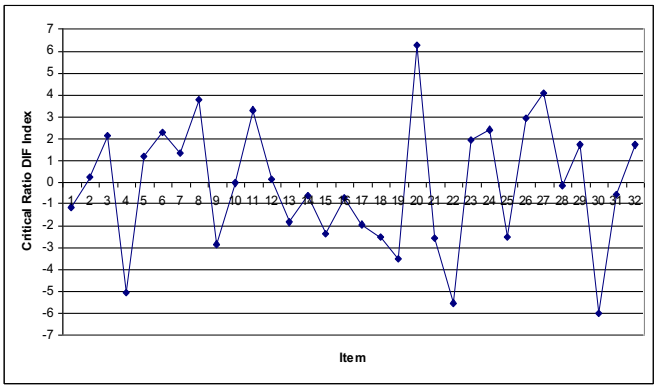

Figure 4: Presentation of the Complete Set of BILOG DIF Indices

Conclusions

The aim of the current investigation was to explore the performance of DIF indices which are produced by Mantel-Haenszel procedure, standardized proportion difference procedure, and critical ratio test (BILOG-MG). The MH and SPD are non-parametric procedures, while the critical ratio test is a parametric one and is based on IRT. These procedures have been extensively studied and reported in the literature. These procedures share a common characteristic that they look at item difficulty differences in the respective groups. In MH, a table of test-taker data is constructed based on item performance, group membership, and score on an overall proficiency measure. The standardization procedure is based on the comparison of the item score in different groups once the level of the ability has been matched. In BILOG-MG, item difficulty difference among the groups under consideration is adjusted and divided by their respective standard errors.

In the present study, an empirical data set was used to analyze the characteristics of each of these DIF detection procedures. Following conclusions have been drawn from the current investigation. Both MH and SPD performed much alike in flagging and classification of DIF items. They also provide a magnitude of DIF, effect size, besides the statistical significance of the test. The effect size facilitates the classification of DIF items with respect to the level of DIF and further necessary actions. In BILOG-MG, no such measure is produced, and flagging criterion is only based on the significance of the critical ratio test. In the absence of effect size measure, inflation to flag items as exhibiting DIF while they are practically trivial can be happen. This phenomenon has been demonstrated using an empirical example. In applied settings MH and SPD are more robust and useful and can be used assertively. From test development perspective, MH should be used in operation for screening unfair items. The user can choose the most appropriate strategy and determine the matching parameters, that is to say, the number of observations or the percentage of the sample that must be considered for the analyses when defining the matching criteria levels. Some empirical guidelines are discussed under the heading of levels of matching variable. For instance for MH, the present investigation has considered thick matching (Min. freq of 5) as an alternative of thin matching (default) and found them similar in functioning. For SPD, the number of respondents in the focal group was chosen as a weighting factor and found to produce analogous results with MH. Finally, when at least 30% of the items on a test are showing DIF then a two-stage DIF strategy should be employed (Zenisky, Robin & Hambleton, 2009).

References

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269-279.

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. En PW Holland and H. Wainer (Eds.), *Differential Item Functioning*, New Jersey: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing the unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Donoghe, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131-154.

Guilera, G., Gómez-Benito, J., & Hidalgo, M. D. (2009). Scientific production on the Mantel-Hanszel procedure as a way of detecting DIF. *Psicothema, 21,* 492- 498.

Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P.

Peterson, E., Baker, & McGaw, B. (Eds.), *International Encyclopedia of Education (3rd edition)*. USA: Elsevier - Science & Technology.

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test Purification and the Evaluation of Differential Item Functioning with Multinomial Logistic Regression. *European Journal of Psychological Assessment, 19*, 1-11.

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

Holland, P W & Wainer, H (Eds.) (1993). Differential item functioning. Lawrence Erlbaum.

Kim S.H., and Cohen, A.S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis [Computer Program] University of Wisconsin-Madison.

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

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement 17*, 297-334.

Muraki, E., & Engelhard, G. (1989, April). *Examining differential item functioning with BIMAIN*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items. Chicago, IL: Scientific Software International.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*, 49-62.

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