DOCUMENT RESUME

ED 307 280 TM 013 212

AUTHOR Engelhard, George, Jr.; And Others
TITLE An Empirical Comparison of Mantel-Haenszel and Rasch Procedures for Studying Differential Item Functioning on Teacher Certification Tests.
PUB DATE 12 Apr 89
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Administration; *Black Students; Comparative Analysis; Early Childhood Education; Elementary Secondary Education; *Item Analysis; Latent Trait Theory; Licensing Examinations (Professions); Racial Bias; Racial Differences; Research Methodology; State Programs; Supervision; *Teacher Certification; *Test Bias; Testing Programs; *White Students
IDENTIFIERS Differential Item Performance; Georgia Teacher Certification Testing Program; *Mantel Haenszel Procedure; *Rasch Model; Teacher Competency Testing

ABSTRACT

The agreement between Mantel-Haenszel and Rasch procedures for identifying differential item functioning (DIF) on teacher certification tests was studied. Two specific research questions were addressed: (1) whether the Mantel-Haenszel and Rasch procedures identify the same items as functioning differently; and (2) how consistently each method identifies items with DIF over administrations. The sample included all black and white examinees who took one of the Georgia Teacher Certification Tests during the December (1987), March (1988), or June (1988) administrations. Item data from these three administrations within the content fields of early childhood (n=1,344; n=1,291; and n=1,023, respectively), middle childhood (n=1,009; n=845; and n=785, respectively), and administration and supervision (n=220; n=216; and n=252, respectively) were used in the analyses, and the differential performance of black and white examinees on these items was examined. The agreement between the two procedures was fairly high within the three administrations, but it dropped significantly when common items were examined across the administrations. The reliability of each procedure was also examined. The data suggest that the Rasch procedure is more consistent in identifying items with DIF than is the Mantel-Haenszel procedure. Further, the data suggest that quantitative indices of DIF are preferable to categorical indices for both procedures. Promising areas for future research on DIF are discussed, and the implications of the findings for theory and practice within the context of teacher certification tests are presented. Six tables present study data. (SLD)

# AN EMPIRICAL COMPARISON OF MANTEL-HAENSZEL AND RASCH PROCEDURES FOR STUDYING DIFFERENTIAL ITEM FUNCTIONING ON TEACHER CERTIFICATION TESTS

George Engelhard, Jr.

Emory University

David Anderson and Stephen Gabrielson

Georgia Assessment Project

Georgia State University

Address: Professor George Engelhard, Jr.
Emory University
Division of Educational Studies
210 Fishburne Building
Atlanta, GA 30322

Running head: MH AND RASCH PROCEDURES

[mhrasch - Paper presented at the annual meeting of the American Educational Research Association Meeting, March 1989]

April 12, 1989

Abstract

This study examines the agreement between Mantel-Haenszel and Rasch procedures for identifying differential item functioning (DIF) on teacher certification tests. Item data from three administrations of teacher certification tests within the content fields of Early Childhood, Middle Childhood, and Administration and Supervision are used in the analyses, and the differential performance of black and white examinees on these items is examined. The agreement between the two procedures is fairly high within the three administrations, but drops significantly when common items are examined across administrations. The reliability of each procedure was also examined, and the data suggest that the Rasch procedure is more consistent in identifying items with DIF than the Mantel-Haenszel procedure. Further, the data suggest that quantitative indices of DIF are preferable to categorical indices for both procedures. Promising areas for future research on differential item functioning are discussed, and the implications of the findings for theory and practice within the context of teacher certification tests are presented.

# AN EMPIRICAL COMPARISON OF MANTEL-HAENSZEL AND RASCH PROCEDURES FOR STUDYING DIFFERENTIAL ITEM FUNCTIONING ON TEACHER CERTIFICATION TESTS

There have been a wide variety of methods proposed for examining item bias (Berk, 1982) or what has come to be called differential item functioning (DIF). These methods have ranged from chi-square methods (Scheuneman, 1979) to methods based on item response theory (Wright, Mead & Draba, 1976; Lord, 1980). Recently, Holland and Thayer (1983) have proposed another method for examining DIF based on methods originally developed by Mantel and Haenszel (1959). This approach shares many of the characteristics of the earlier chi-square methods, but also provides an empirical estimate of the direction and size of subgroup differences on each item.

Linacre and Wright (1986) have highlighted the major similarities and differences between the Mantel-Haenszel (MH) procedure and the use of the Rasch measurement model to examine differential item functioning. The two approaches share many of the same assumptions with major differences related to the issue of how to form score groups for the MH procedure.

This study was conducted in order to increase our knowledge about the correspondence between these two methods using empirical data within the context of teacher certification tests. This

study differs from much of the previous research on differential item functioning in several ways. First, this study focuses on an empirical comparison of the two procedures rather than the analytical comparisons which have been presented earlier by Holland and Thayer (1988) and Linacre and Wright (1986). Any discrepancies identified between the two procedures using empirical data should contribute to our knowledge about both procedures, and their use to flag items which appear to function differently for black and white examinees on teacher certification tests. Second, much of the previous empirical research on DIF using the MH procedure has been based on student achievement data (Raju, Bode & Larsen, 1989; Perlman, et al., 1988; Schulz, Perlman, Rice & Wright, 1989), and it is important to examine the utility of this procedure within the context of teacher certification testing. Finally, one of the practical problems encountered on teacher certification tests is that for some content fields the sample sizes are fairly small, and this study provides evidence regarding how these two procedures perform in these low-incidence fields.

Purpose

The purpose of this study is to compare the Mantel-Haenszel (MH) procedure for examining differential item functioning (DIF) with methods based on the Rasch measurement model. Two specific research questions are addressed: (1) Do the MH and Rasch

procedures identify the same items as functioning differently? (2) How consistently does each method identify items with DIF over administrations? This study also explores the influence of sample size on the two procedures. Two of the content fields examined here (Early Childhood and Middle Childhood) have adequate sample sizes, while one of the content fields (Administration & Supervision) does not meet the minimum sample size requirements of about 100 to 200 examinees recommended for Rasch measurement. This low-incidence field was included in order to gain insight into how these two procedures compare under this condition. Given the close correspondence between the MH and Rasch procedures (Holland & Thayer, 1988; Linacre & Wright, 1986), it is expected that the empirical analyses will yield a high degree of consistency with both procedures flagging similar items for further study by a bias review committee. It is also expected that small sample sizes will decrease the reliability estimates for both procedures and therefore attenuate the estimates of agreement between the two procedures.

## Method

### Subjects

The sample for this study includes all of the black and white examinees who took one of the Georgia Teacher Certification Tests (TCTs) during the December (1987), March (1988) or June (1988)

administrations. The specific fields examined are Early Childhood (n = 1,344, 1,291, and 1,023), Middle Childhood (n = 1,009, 845, and 785), and Administration & Supervision (n = 220, 216, and 252). The description of the samples is presented in Table 1.

Insert Table 1 about here

Instruments

The TCTs are field-specific tests of the content-related knowledge of teachers, and consist of 150 to 250 items per form. These items represent objectives which were identified as being essential skills for minimally competent teachers. All new teachers, as well as teachers seeking re-certification, are required to take the appropriate TCT before they can teach in Georgia. Within each content field, a different test form was given for each of the three administrations. The number of common items between administrations ranges from 44 to 79. The items analyzed for this study include operational and field test items. Estimates of the reliabilities (KR20s) of these tests ranged from .879 to .921, and the summary statistics for each administration are presented in Table 1.

Procedures

The log of the Mantel-Haenszel odds ratio statistic was

calculated for all of the items, and used as a quantitative index of DIF for black and white examinees. The total distribution of raw scores was examined, and divided into six score groups with approximately 16.7 percent of the examinees in each score group. The studied item was included in the formation of the score groups, and in the calculation of the DIF indices. A categorical index was also created with three categories (favor blacks, no difference and favor whites) using the MH chi-square to determine statistical significance (alpha = .05), and the MH odds ratio statistic to determine the direction of differential item functioning. Holland and Thayer (1988) should be consulted for further details on the use of the Mantel-Haenszel procedure to identify differential item functioning.

The $\underline{t}$ statistic recommended by Wright and Stone (1979) was used as the quantitative index of DIF for the Rasch model. A categorical index of bias for the Rasch model was also created. Items with $\underline{t}$ statistics above 2.00 were classified as favoring whites, values below -2.00 were classified as favoring blacks, and the other items were classified in the no difference category.

Pearson and Spearman correlations were calculated for each field using the log of the MH odds ratio statistics and the Rasch $\underline{t}$ statistics. These correlations provide an indicator of the overall agreement between the two procedures. Agreements between the

categorical indices of DIF were also examined using percent agreements and kappa statistics (Cohen, 1960; Fleiss, 1981); the interpretations of kappa statistics are based on the categories proposed by Landis and Koch (1977). In addition to examining the agreement between the two methods within administration, summary statistics were also calculated between administrations. For example, the Rasch t statistic for common items in the December administration can be correlated with the MH statistics from the June administration; conversely, the MH statistics from December administration can be correlated with the Rasch t statistics from the June administration.

The reliabilities for each procedure were also estimated based on common items administered at different times. Pearson and Spearman correlations were calculated for the quantitative indices of DIF, while percent agreements and kappa statistics were calculated for the categorical indices.

## Results

The summary statistics for the black and white examinees by time of administration and content field are presented in Table 1. The raw score means for the white examinees are consistently higher than the raw score means for the black examinees. The results for the quantitative and categorical indices of DIF are reported separately below.

Quantitative Indices of DIF

Table 2 presents the Pearson correlations between the two

Insert Table 2 about here

quantitative indices of DIF by time of administration and content field. The Pearson and Spearman rank order correlations are virtually identical, and therefore only the Pearson correlations are reported here. Four items were deleted from the Administration and Supervision test because all the examinees within one of the comparison groups answered these items correctly. The agreement within administrations appears to be quite good with a median correlation of .774 across time of administration and content field. The correlation for the June administration of the Early Childhood test was somewhat lower than expected, $\underline{r}$ = .554. An examination of the scatterplot indicates that this is due to several items with relatively large Rasch $\underline{t}$ statistics which did not have a correspondingly high log odds ratios.

The Pearson correlations between administrations for the two quantitative indices of DIF are also presented in Table 2. These correlations are consistently smaller with a median correlation across time of administration and content field of .547 as compared to the median correlation within administrations of .774.

The size of the correlations do appear to be related to content field which probably reflects the influence of sample size; the results for the content field of Administration and Supervision (Mdn = .386), which had the smallest sample sizes, are consistently smaller as compared to the content fields of Early Childhood (Mdn = .555) and Middle Childhood (Mdn = .622).

Table 3 shows the Pearson correlations across time of administration for each procedure and content field which provides evidence related to the reliability of each procedure. For the Rasch procedure, the reliability estimates are quite good for Early Childhood (Mdn = .891) and Middle Childhood (Mdn = .904), while the results for Administration and Supervision are clearly lower (Mdn = .512). This trend across content fields also appears to be related to the small sample sizes obtained for Administration and Supervision. The reliability estimates for the MH procedure are presented in the bottom of Table 3.

---

Insert Table 3 about here

---

Pearson correlations are consistently smaller for the MH procedure (Mdn = .497) as compared to the Rasch procedure (Mdn = .860). The trend across content fields for the MH procedure is similar to the trend for the Rasch procedure with both Early Childhood (Mdn =

.482) and Middle Childhood (Mdn = .633) exhibiting higher Pearson correlations than those for Administration and Supervision (Mdn = .294).

Categorical Indices of DIF

Turning now to the results for the categorical indices of DIF, the number of items identified by each procedure as having significant DIF are presented in Table 4. The results of this

Insert Table 4 about here

analysis indicate that the Rasch procedure consistently flags more items than the MH procedure. For the Early Childhood tests, a higher percent of items with DIF were identified by the Rasch procedure (Mdn = 53.3) than the MH procedure (Mdn = 23.4). The results were similar for Middle Childhood with the Rasch procedure (Mdn = 47.5) flagging more items as compared to the MH procedure (Mdn = 21.2). The overall numbers of items flagged by both procedures are smaller for Administration and Supervision; the Rasch procedure identified a median percent of 13.6 items with DIF, while the MH procedure flagged a median percent of 6.4 items.

The number of items identified with DIF by each procedure appears to be related to sample size. As sample size decreases, both methods flag fewer items. The influence of sample size on

differences in the number of items identified with DIF by each method has an effect on the agreement indices which are reported below. It should also be noted that the percentages of items identified in the favor blacks and favor whites categories are different for the Rasch and MH procedures; the Rasch procedure tends to flag comparable percentages of items favoring blacks and whites, while the MH procedure appears to flag more items in the favor blacks categories than the favor whites categories.

The agreement between the categorical indices of DIF (favor blacks, no difference, favor whites) obtained from the Rasch and MH procedures are presented in Table 5.

---

Insert Table 5 about here

---

The median percent agreement within administration is 70.7 and a median kappa statistic of .418 was found over time of administration and content field. For the between administration results, the median percent agreement is 67.0, while the median kappa statistic is .228. Based on the kappa statistics, the data suggest that the agreement for the two procedures is higher within administrations as compared to between administrations; the agreement is moderate within administrations, while fair agreement was observed between administrations.

An examination of Table 5 seems to suggest that the agreement between the two procedures within administrations tends to increase as the sample sizes become smaller. Although this seems to contradict the results based on the quantitative indices of DIF, the explanation for this finding depends on the number of items identified with statistically significant DIF by each procedure. As pointed out earlier, a decrease in sample size leads to fewer items identified with DIF by both procedures and therefore a higher agreement between the Mantel-Haenszel and Rasch procedures is obtained.

Evidence regarding the reliabilities of the categorical indices are presented in Table 6. The median percent agreements for the Rasch (Mdn = 79.6) and the MH procedures (79.6) are equivalent. The kappa statistics are higher

---

Insert Table 6 about here

---

for the Rasch procedure (Mdn = .626) which reflects substantial agreement as compared to the MH procedure (Mdn = .231) which suggests fair agreement over time of administration. The kappa statistics for the Rasch model suggest that reliability decreases as a function of content field; the smallest kappa statistics are found for the Administration and Supervision tests. The kappa

statistics for the reliability of the MH procedure do not appear to be systematically related to content field.

## Discussion

The results of this study suggest that the agreement between the Mantel-Haenszel and Rasch procedures for examining differential item functioning (DIF) is generally quite good within the three administrations of the teacher certification tests examined here. The agreement is lower when common items are examined across administrations. The data also suggest that the Rasch procedure is more reliable than the Mantel-Haenszel procedure. The results are similar for the quantitative and categorical indices of DIF, although other factors, such as sample size and the power of the statistical criterion used to form the categories, make the results less straightforward for the categorical indices.

Before discussing the implications of this study, there are several important issues which should be pointed out. First, the results of this study and other comparison studies may be affected by the methods used to form the score groups with the MH procedure. Further research is needed to provide clear decision rules for the creation of score groups. Some of the factors which need to be explored are the influence of number of score groups, range of scores, distribution of scores within each subgroup, method used to form score groups (e.g., equal percentile groups, fixed score

ranges), the metric of the test scores (e.g., raw score scale, logistic scale), and finally whether or not the studied item is included. The score groups provide a control for differences between subgroups on the variable of interest, and different methods of creating score groups may affect the results of the MH procedure.

Another important issue is related to the influence of sample size. This study examined two content fields where the sample sizes were large enough to justify the use of both procedures, and a third field (Administration and Supervision) where minimum recommended sample sizes of about 100 to 200 for Rasch measurement are clearly not met. Chi-square methods are generally considered more appropriate when the sample sizes are small, and the assumptions of measurement models based on item response theory are not justified. The results of this study suggest that even with the very small sample sizes obtained within the content field of Administration and Supervision, the agreement within administration was quite high between the two procedures when the quantitative indices of DIF are used. Further, it appears that even under these conditions where the minimal sample size requirements are not met, the Rasch procedure still exhibited higher reliability than the MH procedure. It should be noted that the low-incidence content field (Administration and Supervision) included more items than the tests

in the other two fields. Future research should examine the extent to which the number of items may be a factor in explaining the relationships examined here.

Given the close theoretical correspondence between the MH and Rasch procedures, another important issue which needs to be explored further is why the agreement between the quantitative indices of DIF obtained from two procedures was not even higher. One factor is related to the reliabilities of the two procedures which may attenuate the estimates of the agreement. Another factor which was mentioned above is the influence of the method used to create the score groups within the MH procedure.

Finally, it should be noted that the Rasch procedure consistently identified more items with DIF than the MH procedure when the categorical indices of DIF were used. There are several possible explanations for these differences. First, the number of score groups may affect the power of the chi-square statistics to detect differences. Decreasing the number of score groups may lead to more items identified with DIF by the MH procedure as found by Raju, Bode and Larsen (1989). Second, the inclusion of items with significant DIF in the total scores used to create the score groups may also affect the power of the MH procedure because an adequate control for group differences may not be obtained. For example, previous research (Raju, Bode & Larsen, 1989) on the inclusion of

the studied item (as done here) suggests that fewer items may be flagged with DIF. A third factor is that the statistical tests used to form the categories within each procedure may not be examining equivalent hypotheses. Further research is needed on the comparability of the Rasch t statistic and the MH chi-square within the context of research on DIF. Finally, no attempt is made in this study to control for Type I errors. Repeated comparisons between groups on multiple items within a test are not likely to be independent and additional research is needed on this problem. The data reported here also indicates that the MH procedure identifies a greater percentage of items as favoring blacks, while the Rasch procedure appears to flag comparable percentages of favor black and favor white items. Additional research is needed to explore this issue.

Given these issues, there are several important implications of this study for practice. First, test developers need to be aware of the differences in the reliabilities of the methods used to examine differential item functioning. Since a variety of factors can influence the reliability of the method, test developers should explore the reliability of several indices of DIF for their particular tests and examinee populations.

A second implication is that quantitative indices of DIF should be preferred over categorical indices. The agreement

between the MH and Rasch procedures is lower when categorical indices (favor blacks, no difference, favor whites) were used in the analyses. This is due to a loss of information when the metric of the DIF index is categorized and also to the somewhat arbitrary nature of the statistical criteria used to form these categories. Since empirical indices of DIF are generally used by judges within bias review committees to flag items which are potentially biased, a quantitative index which includes an estimate of the size and direction of DIF is more informative and reliable than a categorical index.

In summary, although additional research is still needed on the similarities and differences between the Mantel-Haenszel and Rasch procedures, the results of this study suggest that two procedures are quite similar when the quantitative indices are used, although there are still questions which must be addressed regarding the formation of score groups with the MH procedure. The data suggest that the Rasch procedure is more reliable than the MH procedure. Further, this study also raises a number of cautions related to the use of categorical indices based on either procedure. Further research comparing the two methods with both real and simulated data with known levels of DIF would contribute to our understanding of both procedures.

References

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. Second Edition. New York: John Wiley & Sons.

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

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Linacre, J. M. & Wright, B. D. (1986). Item bias: Mantel-Haenszel and the Rasch model. Memorandum No. 39. Chicago, IL: MESA Psychometric Laboratory, Department of Education, The University of Chicago.

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

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.

Perlman, C. L., Bezruczko, N., Junker, L. K., Reynolds, A.J., Rice, W. K. & Schulz, E. M. (1988, April). Investigating the stability of four methods for estimating item bias. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

Scheuneman, J. D. (1979). A method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Schulz, E. M., Perlman, C., Rice, W. K. & Wright, B. D. (1989). An empirical comparison of Rasch and Mantel-Haenszel Procedures for Assessing Item Bias. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wright, B. D., Mead, R. & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. Research Memorandum Number 22. Chicago: Statistical Laboratory, Department of Education, The University of Chicago.

Wright, B. D. & Stone, M. H. (1979). Best test design. Chicago, IL: MESA Press.

Table 1

Summary data

| | Time | N | M | SD | KR-20 | Items |
|---|---|---|---|---|---|---|
| Early Childhood | | | | | | |
| White | December | 1,084 | 119.4 | 13.5 | .892 | 150 |
| Black | December | 260 | 94.3 | 16.4 | .895 | 150 |
| White | March | 1,061 | 121.5 | 13.6 | .899 | 150 |
| Black | March | 230 | 94.8 | 18.1 | .913 | 150 |
| White | June | 779 | 128.1 | 14.2 | .899 | 160 |
| Black | June | 244 | 97.2 | 17.0 | .894 | 160 |
| Middle Childhood | | | | | | |
| White | December | 807 | 117.3 | 13.4 | .886 | 150 |
| Black | December | 202 | 96.2 | 15.8 | .889 | 150 |
| White | March | 592 | 118.7 | 14.6 | .906 | 150 |
| Black | March | 153 | 92.5 | 16.4 | .895 | 150 |
| White | June | 627 | 123.4 | 14.9 | .899 | 160 |
| Black | June | 158 | 92.7 | 17.8 | .903 | 160 |
| Administration/ Supervision | | | | | | |
| White | December | 157 | 200.8 | 16.2 | .879 | 250 |
| Black | December | 63 | 179.3 | 22.5 | .921 | 250 |
| White | March | 170 | 199.8 | 16.3 | .887 | 250 |
| Black | March | 46 | 179.4 | 22.2 | .919 | 250 |
| White | June | 188 | 202.2 | 16.0 | .879 | 250 |
| Black | June | 64 | 181.3 | 22.0 | .919 | 250 |

Table 2

Pearson Correlations Between Mantel-Haenszel (MH) and Rasch (R) Quantitative Indices of Differential Item Functioning

| | Early Childhood | | Middle Childhood | | Administration/ Supervision | |
|---|---|---|---|---|---|---|
| | R | Items | R | Items | R | Items |
| Within administrations | | | | | | |
| R(D), MH(D) | .774* | 150 | .856* | 150 | .834* | 249 |
| R(M), MH(M) | .746* | 150 | .731* | 150 | .776* | 249 |
| R(J), MH(J) | .554* | 160 | .715* | 160 | .800* | 248 |
| Between administrations | | | | | | |
| MH(D), R(M) | .575* | 50 | .622* | 50 | .391* | 79 |
| MH(D), R(J) | .661* | 50 | .701* | 48 | .390* | 79 |
| MH(M), R(J) | .535* | 54 | .621* | 44 | .382* | 75 |
| R(D), MH(M) | .741* | 50 | .559* | 50 | .419* | 79 |
| R(D), MH(J) | .529* | 50 | .660* | 48 | .376* | 78 |
| R(M), MH(J) | .467* | 54 | .593* | 44 | .193 | 75 |

* $p < .05$

Note. Similar results were obtained using the Spearman rank order correlation; December (D), March (M) and June (J) administrations.

Table 3

Pearson Correlations of Mantel-Haenszel and Rasch Quantitative Indices of Differential Item Functioning Over Time

| | Early Childhood | | Middle Childhood | | Administration/ Supervision | |
|---|---|---|---|---|---|---|
| | R | Items | R | Items | R | Items |
| Rasch | | | | | | |
| R(D), R(M) | .897* | 50 | .824* | 50 | .512* | 79 |
| R(D), R(J) | .860* | 50 | .917* | 48 | .626* | 79 |
| R(M), R(J) | .891* | 54 | .904* | 44 | .505* | 75 |
| Mantel-Haenszel | | | | | | |
| MH(D), MH(M) | .666* | 50 | .597* | 50 | .497* | 79 |
| MH(D), MH(J) | .476* | 50 | .633* | 48 | .294* | 78 |
| MH(M), MH(J) | .482* | 54 | .756* | 44 | .165 | 75 |

* $p < .05$

Note. Similar results were obtained using the Spearman rank order correlation; December (D), March (M) and June (J) administrations.

Table 4

Percent of Items with Significant Differential Item Functioning for the Rasch and Mantel-Haenszel Procedures

| | Favor Blacks | | No Difference | | Favor Whites | |
|---|---|---|---|---|---|---|
| | Rasch | MH | Rasch | MH | Rasch | MH |
| Early Childhood | | | | | | |
| December (150) | 28.0 | 10.7 | 46.7 | 69.3 | 25.3 | 20.0 |
| March (150) | 30.7 | 6.7 | 45.3 | 76.6 | 24.0 | 16.7 |
| June (160) | 21.2 | 4.4 | 54.4 | 88.7 | 24.4 | 6.9 |
| Middle Childhood | | | | | | |
| December (150) | 26.7 | 15.3 | 48.0 | 65.3 | 25.3 | 19.3 |
| March (150) | 21.3 | 6.7 | 57.3 | 79.3 | 21.3 | 14.0 |
| June (160) | 23.1 | 5.0 | 52.5 | 78.8 | 24.4 | 16.2 |
| Administration/ Supervision | | | | | | |
| December (249) | 11.6 | 6.0 | 77.2 | 83.6 | 11.2 | 10.4 |
| March (249) | 8.0 | 1.6 | 88.4 | 95.6 | 3.6 | 2.8 |
| June (248) | 6.8 | 2.0 | 86.4 | 93.6 | 6.8 | 4.4 |

Note. Number of items are given in parentheses.

Table 5

Percent Agreements and Kappa Statistics for Mantel-Haenszel and Rasch Categorical Indices of Differential Item Functioning

| | Early Childhood | | Middle Childhood | | Administration/ Supervision | |
|---|---|---|---|---|---|---|
| | Agree | Kappa | Agree | Kappa | Agree | Kappa |
| Within administrations | | | | | | |
| R(D), MH(D) | 65.3 | .418* | 70.7 | .508* | 86.4 | .595* |
| R(M), MH(M) | 58.7 | .302* | 72.7 | .454* | 91.2 | .423* |
| R(J), MH(J) | 57.9 | .134* | 61.2 | .276* | 88.8 | .401* |
| Between administrations | | | | | | |
| MH(D), R(M) | 60.0 | .231* | 76.0 | .545* | 79.7 | .201* |
| MH(D), R(J) | 68.0 | .385* | 64.6 | .347* | 81.0 | .326* |
| MH(M), R(J) | 61.1 | .207* | 72.7 | .396* | 85.5 | .100 |
| R(D), MH(M) | 62.2 | .265* | 62.0 | .182 | 79.7 | .224* |
| R(D), MH(J) | 54.0 | .131 | 58.3 | .272* | 74.7 | .168* |
| R(M), MH(J) | 64.8 | .191* | 65.9 | .253* | 86.8 | .181* |

* Value is more than twice the standard error.

Note. Number of items is the same as in Table 2; December (D), March (M) and June (J) administrations.

Table 6

Percent Agreements and Kappa Statistics for Rasch and Mantel-Haenszel Categorical Indices of Differential Item Functioning Over Time

| | Early Childhood | | Middle Childhood | | Administration/ Supervision | |
|---|---|---|---|---|---|---|
| | Agree | Kappa | Agree | Kappa | Agree | Kappa |
| Rasch | | | | | | |
| R(D), R(M) | 78.0 | .633* | 82.0 | .672* | 81.0 | .337* |
| R(D), R(J) | 72.0 | .538* | 81.2 | .693* | 77.2 | .317* |
| R(M), R(J) | 79.6 | .642* | 79.5 | .326* | 85.5 | .294* |
| Mantel-Haenszel | | | | | | |
| MH(D), MH(M) | 74.0 | .241 | 70.0 | .292* | 81.0 | .158* |
| MH(D), MH(J) | 78.0 | .328* | 60.4 | .110 | 81.0 | .231* |
| MH(M), MH(J) | 79.6 | .185 | 86.3 | .446* | 90.8 | .183 |

* Value is more than twice the standard error.

Note. Number of items is the same as in Table 3; December (D), March (M) and June (J) administrations.