DOCUMENT RESUME

ED 390 887 TM 024 168

AUTHOR	Allen, Nancy L.; Donoghue, John R.
TITLE	Application of the Mantel-Haenszel Procedure to Complex Samples of Items.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	ETS-RR-95-4
PUB DATE	Jan 95
CONTRACT	R89046001
NOTE	51p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, 1991).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Computer Assisted Testing; Difficulty Level; Elementary Secondary Education; *Identification; *Item Bias; Item Response Theory; Models; Monte Carlo Methods; National Surveys; *Sampling; Test Construction; *Test Items
IDENTIFIERS	Balanced Incomplete Block Spiralling; *Mantel Haenszel Procedure; National Assessment of Educational Progress; Three Parameter Model

ABSTRACT

This Monte Carlo study examined the effect of complex sampling of items on the measurement of differential item functioning (DIF) using the Mantel-Haenszel procedure. Data were generated using a three-parameter logistic item response theory model according to the balanced incomplete block (BIB) design used in the National Assessment of Educational Progress. The length of each block of items and the number of DIF items in the matching variable were varied, as were the difficulty, discrimination, and presence of DIF in the studied item. Block, booklet, pooled booklet, and extra-information analyses were compared to a complete data analysis using the transformed log-odds on the delta scale. The pooled booklet approach is recommended for use when items are selected for examinees according to a BIB design. This study has implications for DIF analyses of other complex samples of items, such as computer administered testing or another complex assessment design. (Contains 14 tables and 20 references.) (Author/SLD)

RR-95-4

RESEARCH REPORT

APPLICATION OF THE MANTEL-HAENSZEL PROCEDURE TO COMPLEX SAMPLES OF ITEMS

NANCY L. ALLEN
JOHN R. DONOGHUE

Educational Testing Service
Princeton, New Jersey
January 1995

Application of the Mantel-Haenszel Procedure to Complex Samples of Items

Nancy L. Allen and John R. Donoghue
Educational Testing Service

A version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1991. The work reported herein was supported under the National Assessment of Educational Progress (Contract No. R89046001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The authors would like to thank John Mazzeo and Rebecca Zwick for their numerous helpful comments and suggestions.

Abstract

This Monte Carlo study examined the effect of complex sampling of items on the measurement of differential item functioning (DIF) using the Mantel-Haenszel procedure. Data were generated using a 3PL IRT model according to the balanced incomplete block design used in the National Assessment of Educational Progress (NAEP). The length of each block of items and the number of DIF items in the matching variable were varied, as was the difficulty, discrimination, and presence of DIF in the studied item. Block, booklet, pooled booklet and extra-information analyses were compared to a complete data analysis using the transformed log-odds on the delta scale. The pooled booklet approach is recommended for use when items are selected for examinees according to a BIB design. This study has implications for DIF analyses of other complex samples of items, such as computer administered testing or another complex assessment design.

Application of the Mantel-Haenszel Procedure to Complex Samples of Items

DIF studies compare the relative performance of the subgroup of interest (the focal group) to that of a comparison or reference group. The Mantel-Haenszel (M-H) procedure (Mantel & Haenszel, 1959) was introduced by Holland and Thayer (1988) to identify items that function differently for the two subgroups. Such items are said to have DIF (differential item functioning). Typically, the M-H procedure has been used with standardized tests presented in traditional formats, i.e., each examinee takes the same collection of items. Even if several forms of a test were administered at the same testing session, forms were usually analyzed separately, even if they have some common items.

Currently, two trends in testing are moving away from traditional testing formats and toward the increasing use of complex sampling of which items are administered to an examinee. The first trend is the increasing use of computers in the administration of tests. Some tests that are presently administered by computer use testlets as the basic grouping of items administered to examinees. For instance, in the prototype of the National Council of Architectural Registration Boards (NCARB) test described by Lewis and Sheehan (1990) and Wainer and Lewis (1990), examinees are presented with testlets of 10-20 items. After the first one or two testlets are administered, a decision to continue or conclude testing is made. If testing is continued, another testlet of 10-20 items is selected randomly and administered.

The other trend away from traditional test formats is the increasing use of the results of large-scale assessments in political decision-making, such as the current debate on school accountability. In large-scale assessments, only results for groups are required. Thus, complex sampling of items is used to ensure the coverage of a large universe of items while simultaneously limiting test time for individual examinees. An example of a complex item sampling plan is the matrix sampling procedure used by the California Achievement Program (Bock & Mislevy, 1981). In this procedure, several booklets containing different items are administered randomly to samples of students. Another complex item sampling procedure is the balanced incomplete block (BIB) des .n used in the National Assessment for Educational Progress (NAEP) (Johnson & Allen, 1992). Table 1 displays the BIB design used in the 1990 NAEP Mathematics Assessment. Items are grouped into seven separately timed groups, termed "blocks." Individual NAEP blocks are usually small, and may contain as few as eight items. These blocks are then combined into seven test booklets, consisting of three blocks each. The design is organized so that each block appears in each position (first, second or third) within a booklet, and each pair of blocks appears together once. As in the matrix sanpling procedure, each booklet is administered to a random sample of students.

Insert Table 1 about here

As with tests presented in traditional formats, DIF information is of interest when items are sampl^d in complex ways. This information can contribute to the appropriate use of particular items in current analyses of data and to the development of more appropriate items for future assessments. However, standard approaches to the application of the M-H procedure can be problematic when applied to data that has been collected according to complex designs. Complex sampling of items, such as the NAEP BIB design, results in relatively sparse data for individual items. This raises questions about the most appropriate method of forming the M-H matching variable for complex samples of items. This study compares several methods of applying the M-H procedure to data generated according to the seven block/seven booklet BIB design used in NAEP. The chief question addressed is: What should the matching variable be when data are collected according to the NAEP BIB design?

The MH DIF Procedure

The M-H procedure matches the reference and focal groups on some measure of performance. In usual DIF applications of M-H to tests with traditional formats, this "matching variable" is the total score on the test. For each of the K levels of the matching variable, M-H forms a 2 X 2 table, which is shown in Table 2. T_k is the total number of examinees at level k, n_{Rk} and n_{Fk} are the number of reference and focal group members, m_{1k} is the number of examinees who answered the studied item correctly, and m_{0k} is the number who missed the item.

Insert Table 2 about here

In applying the M-H procedure, it is assumed that the odds-ratio α is constant across the K levels of the matching variable. The M-H statistic $\hat{\alpha}_{MH}$ estimates a pooled odds-ratio under this assumption:

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{K} A_k D_k / T_k}{\sum_{k=1}^{K} B_k C_k / T_k} \quad . \tag{1}$$

The MH statistic is often transformed, in psychometric applications, to:

$$\hat{\Delta}_{MH} = -2.35 \cdot \log_e(\hat{\alpha}_{MH}) \quad . \tag{2}$$

This transformation makes the measure negative for items which are harder (conditional on values of the matching variable) for the focal group, and puts $\hat{\Delta}_{MH}$ on the "delta-scale" used at ETS to measure item difficulty. The estimated standard error (Holland & Thayer, 1988) is:

$$se(\hat{\Delta}_{MH}) = 2.35 \cdot \sqrt{Var(\log_e(\hat{\alpha}_{MH}))} \tag{3}$$

where

$$Var(\log_e(\hat{\alpha}_{MH})) = \frac{\sum_{k=1}^{K} \frac{U_k V_k}{T_k^2}}{2\left(\sum_{k=1}^{K} \frac{A_k D_k}{T_k}\right)^2} \quad (4)$$

$$U_k = (A_k D_k) + \hat{\alpha}_{MH}(B_k C_K)$$

$$V_k = (A_k + D_k) + \hat{\alpha}_{MH}(B_k + C_k) \quad .$$

In addition, a one-degree of freedom χ^2 test is available (see Holland & Thayer, 1988).

Several studies have examined the attributes of the Mantel-Haenszel approach to examining DIF since Holland and Thayer introduced the method for this purpose in 1988, but those studies have largely dealt with traditional testing formats. To date, only three studies have examined differential item functioning with NAEP data. Zwick and Ercikan (1989) looked at DIF for items presented as part of the 1986 NAEP history assessment. They examined items identified as displaying DIF on the basis of racial-ethnic group membership. Unexpectedly, M-H analyses that also incorporated exposure to history material did not reduce the number of items identified as having DIF. This result may be due to the sparseness of the data in the 2 X 2 tables when conditioning takes place on history exposure variables in addition to race/ethnicity.

The second study, by Nelson and Zwick (1989) and later replicated by Zwick and Grima (1990), presented information about the effect of the complex sampling of students used in NAEP on examinations of DIF. These studies indicated that the use of sampling weights can drastically influence the results of DIF analyses. Calculating jackknife standard errors for the Mantel-Haenszel statistics (as is done for other statistics in NAEP) had little influence on the results of DIF analyses, however.

The third study (Zwick & Grima, 1990) was reported in a comprehensive internal document regarding appropriate DIF analyses for NAEP assessments. Among other results, they found that, for real NAEP data, there was no indication that the use of the block analysis produced spurious DIF. Zwick and Grima also found that the booklet analyses produced M-H DIF statistics with larger standard errors than those for block analyses. The larger standard errors could be due to context effects, block position effects, multidimensionality, or increased sampling variability because of reduced sample size.

Mantel-Haenszel Approaches to Complex Samples of Items

As noted above, the total number correct on a test form is usually used as the matching variable. If multiple forms of a test are used, each form is administered to a large number of examinees, and each form is analyzed separately. Complex sampling of items, such as the NAEP BIB design, results in relatively sparse data for individual items. This raises questions about the most appropriate method of forming the matching variable for complex samples such as the BIB design. In the discussion that follows, we will assume that the studied item is contained in Block A. To illustrate the various approaches, Table 3 schematically shows the information available about a given item in Block A from the BIB design portrayed in Table 1[1].

Insert Table 3 about here

In block level matching, the traditional total score matching variable is computed as the sum of the items in Block A. This has the advantage that the M-H statistic for each item are based on three times the number of students administered any particular booklet (e.g., 2000 students per booklet, 6000 per block). However, the block level matching variable may not be sufficiently reliable, because it could be based on as few as eight items in some NAEP subject areas. Donoghue, Holland, and Thayer (1993) indicated that

[1] For ease of illustration, we have assumed that each block contains 10 items. In practice, the blocks of the BIB are not required to be of the same length.

total score matching variables based on fewer than 20 items can adversely influence the M-H statistics. Zwick (1990) corroborated the fact that using the M-H procedures can be problematic when the matching score is unreliable.

The traditional total score matching variable can also be used separately for each booklet. The advantage of booklet level matching is a more reliable matching variable, based on about 24 to 60 items. However, this approach results in multiple measures of DIF for the same item. In the BIB design described above, three M-H statistics (based on Booklets 1, 2, and 3) would be calculated for each item in Block A. The three matching variables for these M-H statistics are represented at the bottom of Table 3. Another potential drawback of booklet level matching is that the number of examinees receiving the same booklet is much smaller than the number of examinees receiving a common block of items. Therefore, the individual M-H DIF statistics calculated at the booklet level will be subject to greater sampling variability than those computed from the larger sample of examinees used for block level matching.

Alternative Methods of Forming the Matching Variable

The M-H procedures were developed in the context of meta-analysis, where the only assumption made about the relationship between the 2 X 2 tables was that the odds-ratio for each of the tables has the same value. In other words, the M-H procedures do not assume any relationship between the levels of the matching variable; it is only assumed that each level has a common odds-ratio. In this study, we examine two alternatives to the block and booklet level M-H analyses that take advantage of this property. Each alternative produces a single M-H statistic for each item while making use of the addition information available in the rest of the booklet.

The first alternative (the pooled booklet approach) pools the information from each of the three booklet M-H analyses for an item. Each of the booklet level analyses for an item are based on a 2 X 2 X k_i frequency table, where k_i is the number of score levels in the matching variable, in this case the number right for the items in booklet i ($i = 1,2,3$). The pooled booklet M-H statistic is based on the 2 X 2 X ($k_1 + k_2 + k_3$) table made by concatenating all of the 2 X 2 tables in the three booklet level analyses. This approach has the advantage of producing one M-H analysis for each item while taking into account the information contained in every booklet in which the item appears. However, it does make an added assumption that the odds-ratio is constant across *all* of the 2 X 2 tables contributing to the analysis, rather than across the k_i 2 X 2 tables for each of the booklets.

A second alternative approach (the extra-information approach) separates the number right score for Block A (p_A) from the number right score for the other two blocks (q_{iA}) in each of the three booklets. Extra-information matching is schematically portrayed in Table 4. For a single booklet, the matching variable is formed by crossing each level of total score on Block A with each level of total score on the other two blocks in the booklet. The results for each booklet are combined, so that the DIF statistics are based on a 2 X 2 : m_A table, where m_A is p_A times ($q_{1A} + q_{2A} + q_{3A}$). In the example in Table 4 where each block contains 10 items, $p_A = 11$ and $q_{iA} = 21$ for each of the three booklets. Therefore, there are $m_A = 11 * 63 = 693$ levels of the matching variable.

Insert Table 4 about here

The extra-information approach makes the assumption that the odds-ratio is constant across all m_A of the 2 X 2 matrices. It also leads to very small cell sizes. It has the advantage, however, of taking into account the total number right for the block containing the studied item and the peripheral information in the complementary blocks in the three appropriate booklets. A similar approach was used by Zwick and Ercikan (1989) to incorporate the effects of educational experiences of students into the matching variable.

Research Questions

The results that are described here address specific research questions. The most important questions focus on the relative effectiveness of methods of forming the matching variable. First, which of the four approaches described above (block level, booklet level, pooled booklet, or extra-information) yield results most like the analysis based on complete data (i.e., the full data matrix that would be obtained if all examinees received all items)? Complete data analysis is not perfect; the number of items that contribute to the score used as a matching variable is finite, as is the number of people in each subgroup. However, the number of items that contribute to the matching variable is large and the number of people in each subgroup is as large as they ordinarily would be for a NAEP assessment. Second, how similar are each of the three booklet analyses to one another and to the pooled booklet analysis? Third, how short can the length of the block be, while still producing useful results using the block approach? Finally, do the results obtained corroborate those of previous studies?

Results of applying the M-H procedure to the analysis of the complete data will be presented first, followed by the results for using the M-H approaches with complex samples of items. This order of presentation will provide a general context for the results for complex samples of items.

Method

This study used Monte Carlo methods to compare the methods of forming the M-H matching variable described above in a controlled setting. Initially, a full data matrix was generated where all examinees received all items. As a baseline, a standard M-H analysis analyzing complete item data, generated as if every examinee received every item, was computed. Then items were deleted from the matrix for individual examinees to fit the portion of the BIB design[2] portrayed in Table 3, and M-H analysis using each of the four methods was applied to the reduced data set.

Simulated test data were generated with a three-parameter logistic (3PL) item response (IRT) model (Birnbaum, 1968). The items were generated to represent dichotomously scored items. No omitted or not-reached items were allowed, so the only missing data was due to the item sampling structure.

Data Simulation and Summarization

Design of the Data Generation

Several independent variables were manipulated. The variables of major interest were whether or not the studied item contained DIF and the M-H approach used in the analysis of the item. In addition, four other variables were varied, due to the known influence of these variables on M-H DIF analyses. The values and levels of the independent variables were selected to reflect the magnitude of values seen in NAEP data. The independent variables were:

Variables determining the BIB design condition

NBLK - Length of each block, including the studied item (3 levels)

1) 10 items
2) 20 items
3) 30 items

11 NDIF - Number of DIF items, other than the studied item, in a block (2 levels)

1) no items
2) 2 items (favoring the reference group: $b_F = b_R + 0.4$)

[2] In NAEP, the samples of students selected for the assessment are also complex samples. In this study, however, only complex sampling of groups of items was examined.

Variables defining the studied item

DIF - DIF in the studied item (2 levels)

1) no DIF: $b_{\underline{F}} = b_{\underline{R}}$
2) DIF favoring the reference group: $b_{\underline{F}} = b_{\underline{R}} + 0.4$

b - Difficulty of the studied item (5 levels)

$b_{\underline{R}} = (-2.0, -1.0, 0.0, 1.0, 2.0)$

a - Discrimination of the studied item (3 levels)

$a_{\underline{k}} = (.5, 1.0, 1.5)$

M-H approach

METHOD (7 levels)

1) Block level analysis
2) Booklet level analysis for booklet 1
3) Booklet level analysis for booklet 2
4) Booklet level analysis for booklet 3
5) Pooled booklet analysis
6) Extra-information analysis
7) Analysis of complete data

In generating the data, the length of each block and number of DIF items, other than the studied item, in a block were fully crossed. These two factors define a BIB design condition. DIF in the studied item, and discrimination and difficulty of the studied item were crossed within each BIB design condition. In design of analysis terms, the data were generated to fit a split plot design, where variables defining the studied item (DIF, *a*, and *b*) and the M-H approach (METHOD) were within-dataset factors and variables determining the BIB design condition were between-dataset factors.

Data Generation and Analysis

To replicate the NAEP setting in a controlled way, sample sizes and item parameters were chosen to approximate values actually observed in NAEP assessments. For the reference group, 5100 abilities were sampled from a normal distribution with mean zero and standard deviation .7. For the focal group, 1050 abilities were sampled from a normal distribution with mean -0.7

and standard deviation .8.

Next, IRT item parameters were selected. For studied items, *a*- and *b*-parameters were determined by the design; 30 studied items were defined by crossing the independent variables DIF, *a*, and *b*. The pseudo-guessing parameter c = 0.2 for all studied items. For items other than studied items, IRT item parameters were sampled randomly (with replacement) from the parameters obtained in the operational calibration of the 1986 age 13 NAEP math trend assessment. These parameter estimates were obtained from *Expanding the New Design: The NAEP 1985-86 Technical Report* (Beaton, 1988). For the NDIF=2 condition, the two items which displayed DIF, other than the studied item, were selected randomly from the nonstudied items within each block.

Responses of all examinees to all items were then generated according to the 3PL model. Each subject's probability of getting each item correct P_{ij} was computed. A uniform [0,1] pseudo-random number v_{ij} was then generated. The item was considered correct if $v_{ij} \leq P_{ij}$, and incorrect otherwise.

For a given BIB design condition, responses to the nonstudied items (each of the NBLK items in Blocks B-G, and items 1 through NBLK-1 of Block A) were generated. In turn, each studied item *j* (*j*=1,...,30) was then considered to be final item of the BIB design (item NBLK of Block A), and DIF statistics were computed for that studied item according to the various approaches. Studied item *j* was then discarded, and the next studied item (*j*+1) was considered to be final item NBLK of Block A. When DIF analyses for each of the 30 studied items had been performed, the dataset was discarded. This process (sampling of examinee abilities, sampling of item parameters for nonstudied items, generation of item responses, and DIF analysis of each studied item in turn) constituted a single replication of that BIB design condition. Fifty replications of each BIB design condition were performed.

The dichotomous item responses were analyzed using the M-H approaches described above. As indicated by previous research (Donoghue, Holland, & Thayer, 1993; Holland & Thayer, 1988; and Zwick, 1990), the matching variable always included the studied item. Four M-H DIF statistics were recorded: the pooled odds-ratio, $\hat{\alpha}_{MH}$; the transformed log-odds, $\hat{\Delta}_{MH}$ and its standard error $SE(\hat{\Delta}_{MH})$; and the Mantel-Haenszel χ^2 (as described by Holland & Thayer, 1988). Primarily, results for $\hat{\Delta}_{MH}$ and its standard error will be presented here.

Data Summarization

As described above, the study included 62,000 (3 X 2 X 2 X 5 X 3 X 7 X 50) observations generated in a complex way. Because the purpose of the study was to examine specific questions and because the amount of data that was generated was large, the methods of analysis were specific to the questions at hand. Results are illustrated by appropriate means for specific BIB design conditions and studied item conditions of the design.

Results

General Results for Using the Mantel-Haenszel Approach to Examine DIF

Results of the Mantel-Haenszel procedure based on the complete data for the case of no items with DIF (other than possibly the studied item) are presented in Table 5. This case was selected to represent the general trends in the $\hat{\Delta}_{MH}$ due to independent variables other than the method of forming the matching variable. As in Donoghue, Holland, and Thayer (1993), there was an overall difference between the $\hat{\Delta}_{MH}$ values for the cases where the studied item was functioning differentially (DIF) as opposed to the cases where the studied item reflected no DIF (NO DIF). The top section of Table 5 demonstrates that NBLK (the number of items in the block) had little influence on the $\hat{\Delta}_{MH}$ value, whether or not the studied item contained DIF. The bottom section of Table 5 shows that, as in the previous study, the effect of the difficulty of the studied item (*b*) was important. As the studied item became more difficult, the $\hat{\Delta}_{MH}$ value became larger; in other words, apparent DIF against the focal group decreased. Larger differences in $\hat{\Delta}_{MH}$ values for extreme *b* values occurred in the condition where the studied item had DIF than in the NO DIF condition.

Insert Table 5 about here

When there were two other items with DIF within each block, $\hat{\Delta}_{MH}$ values were consistently larger than the values when no other items with DIF were in the block. For example, consider the case where the discrimination of the studied item was .5 and its difficulty was -2. When other items with DIF contributed to the matching variable, the mean $\hat{\Delta}_{MH}$ value for the NO DIF case was .09 and the mean for the mean for the DIF cases was -.64. The corresponding means were -.02 and -.73 (see Table 5) when there were no other DIF items, with the other possible exception of the studied item. For studied items which displayed DIF, it appeared that there was less DIF against the focal group when there were two other items with DIF in each block than when there were no other DIF items in the block. Because this was a consistent result, the rest of the results will reflect only the case where no items with DIF were included in addition to the studied item.

Other important effects described by Donoghue, Holland, and Thayer (1993) involved the discrimination of items in the test. In the Donoghue, Holland, and Thayer study, the discrimination of the studied item was the same as the discrimination of every other item in the matching variable, and that value was varied as an independent variable. In this study, the discrimination of the studied item was varied as an independent variable, and the discrimination values of the other items reflected those for actual NAEP items. Therefore, results of this study reflect an interaction between the discrimination of the studied items and the discriminations of the other items. For items with *b*-parameter values where most of the focal and reference group distributions lie ($b = -2, -1, 0$), the pattern of the results in Table 5 is similar to those in the Donoghue, Holland, and Thayer. For these items, as the *a*-parameter of the studied item increased, the difference between mean $\hat{\Delta}_{MH}$ values for the DIF and NO DIF conditions increased. For more difficult items ($b = 1, 2$), however, this was not true.

The mean $\hat{\Delta}_{MH}$ values in Table 5 also indicate how difficult it might be to identify items with DIF under certain extreme conditions. In particular, when the studied item is difficult for both focal and reference group examinees, items with DIF against the focal group may not be identified. On the other hand, when the discrimination of the studied item is large and the difficulty is small, items with NO DIF may be spuriously identified to have DIF against the focal group.

Results for Using Mantel-Haenszel Approaches with Complex Sampling of Items

The mean $\hat{\Delta}_{MH}$ values for each method and DIF or NO DIF in the studied items averaged over the other conditions are listed in Table 6. In examining Table 6, it should be borne in mind that, under the null hypothesis of common odds-ratio and NO DIF, $\hat{\Delta}_{MH}$ is expected to be zero. When the studied item has DIF, $\hat{\Delta}_{MH}$ should be negative. On average, the block analysis mean differed most from the complete data analysis mean when the studied item had NO DIF. The extra-information analysis results differ most from the complete data results when the studied item had DIF. In both the DIF and NO DIF conditions, the booklet and pooled booklet analyses have mean $\hat{\Delta}_{MH}$ values which are close to those for the complete data analysis.

Insert Table 6 about here

The general results for the analysis methods are supported when specific cells in the design are examined. In the NO DIF condition, results for the analysis of the complete data were the least biased; the mean $\hat{\Delta}_{MH}$ value for the analysis of the complete data was closer in absolute value to zero than for the other approaches to forming the matching variable. Table 7 contains mean differences between the $\hat{\Delta}_{MH}$ values for the complete data analyses and the other analysis methods for selected cells in the data generation design when no items other than possibly the studied items contain DIF. In general, the largest differences between the results for the complete data analysis and the other analysis methods occurred when the studied item was very discriminating and very easy ($a = 1.5$, $b = -1$).

Insert Table 7 about here

As expected, booklet level matching exhibited more sampling variability than did the other methods. This effect was seen in three ways. First, for individual replications results of the individual booklets differed noticeably from one another. Second, the standard deviations of values of $\hat{\Delta}_{MH}$ for each of the booklet analyses were higher than for the other methods (see Table 7). Third, this variability was accurately reflected in the larger theoretic standard errors (discussed below) associated with booklet level matching.

Table 8 contains the mean values of the theoretic $SE(\hat{\Delta}_{MH})$ for each of the methods of forming the matching variable. Standard errors were slightly smaller when the studied item contained DIF than when it had NO DIF. When there were two other items with DIF (in addition to possibly the studied item) within each block, standard errors were virtually identical to the condition where there were no other items which contained DIF. The standard errors for each of the booklet analyses were consistently larger than the standard errors for the complete data analyses, while the standard errors of block, and pooled booklet analyses were close to those from the complete data analyses. The extra-information approach yielded the smallest standard errors.

Insert Table 8 about here

Table 9 provides further information about the effect of the method of analysis on the theoretic $SE(\hat{\Delta}_{MH})$. For each cell in the design, the standard deviation of the value of $\hat{\Delta}_{MH}$ for the 50 replications was computed. The mean theoretic $SE(\hat{\Delta}_{MH})$ was also computed for the 50 replications. Table 9 presents results for the ratio of the observed standard deviation of $\hat{\Delta}_{MH}$ obtained for each method to the mean value of $SE(\hat{\Delta}_{MH})$. Values close to one indicate that the average standard error was similar to the observed standard deviation, while values greater than one indicate that the average standard error underestimated the observed variability of $\hat{\Delta}_{MH}$.

Insert Table 9 about here

For the booklet level matching, pooled booklet, and analysis of the complete data, the ratios were near one, indicating $SE(\hat{\Delta}_{MH})$ reflected the variability of $\hat{\Delta}_{MH}$ fairly accurately. Block level matching, however, yielded ratios which were substantially greater than one when the length of the block was either 10 or 20 items. For blocks of 30 items, however, the ratio was close to one indicating that the standard error is fairly accurate. Overall, the $SE(\hat{\Delta}_{MH})$ for the extra-information approach substantially underestimates the variability of the $\hat{\Delta}_{MH}$ values. In addition, the pattern of results for different block lengths was unusual. The ratio was close to one for blocks of 10 items, much greater than one for blocks of 20 items, and closer to one (although still too large) for blocks of 30 items. This pattern was unexpected, and would require additional, more in-depth simulation results to adequately understand it.

The top section of Table 10 shows the relationship between analysis of the complete data and the block level analysis for different numbers of items in the block (NBLK). Averaging over difficulty and discrimination of the studied item, it appears that the effect of increasing the number of items in the matching variable from 10 to 20 to 30 is only marginally important. Although the overall means do not differ, the absolute size of the differences decreases with the block size. This is reflected by the decreasing standard deviations, and in the comparison of specific combinations of *a* and *b* levels (results for the full interaction of DIF X NBLK X *a* X *b* are given in Appendix Table A1). In general the block analysis tends to overestimate the $\hat{\Delta}_{MH}$ values as compared to the complete data analysis when the studied item is not very discriminating ($a = .5$) or when the studied item is very difficult ($b = 1, 2$). The block analysis tends to underestimate the $\hat{\Delta}_{MH}$ values when the studied item is easy and very discriminating ($a = 1, 1.5$; $b = -2, -1, 0$) This is exactly the case where the $\hat{\Delta}_{MH}$ values are underestimated even by the complete data case, indicating that the block analysis results in a larger chance that an item that does not have DIF will be falsely identified as having DIF against the focal group.

Insert Table 10 about here

The relationship between the complete data method and the pooled booklet method for different numbers of items in the matching variable (NBLK) are shown in the bottom portion of Table 10 (results for the full DIF X NBLK X *a* X *b* interaction are given in Appendix Table A2). The magnitude of the differences for the pooled booklet method are always smaller than that of the differences for the block method for the same set of conditions.

To get some sense of the practical importance of the differences between the block and pooled booklet approaches, Table 11 presents information about the classification of the studied items as having DIF using operational definitions in use at Educational Testing Service (ETS). "A" items have $\hat{\Delta}_{MH}$ values which do not significantly differ from 0 and/or are smaller than 1.0 in absolute value. "B" items meet have $\hat{\Delta}_{MH}$ values which significantly differ from 0 and are greater than 1.0 in absolute value, but the $\hat{\Delta}_{MH}$ values of B items are not significantly greater than 1.0 in absolute value and/or are smaller than 1.5 in absolute value. The $\hat{\Delta}_{MH}$ value of a "C" item is both significantly greater than 1.0 in absolute value and greater than 1.5 in absolute value. Finally, the sign attached to the classification reflects the sign of $\hat{\Delta}_{MH}$.

Items classified as C- are items defined as having DIF against the focal group and would be excluded from the item pool for a new test unless content considerations, as operationalized by decisions of an examining committee of subject area specialists, were overwhelming. Items classified as B- are items defined as possibly having DIF against the focal group and would be avoided as much as possible in the development of a new test.

Zwick, Thayer, and Wingersky (1993) examined the relationship between IRT item parameters and the value of $\hat{\Delta}_{MH}$. Based on simulation results, they suggest that, for items with nonzero *c*-parameters, $\hat{\Delta}_{MH}$ is well approximated by $-3*a(b_F-b_R)$. They also report that an item for which $|a(b_F-b_R)| \geq .52$ would be expected to be classified as a C item at least 75% of the time. In the present study, items with DIF had $b_F-b_R = .4$. Thus, in the case where $a=1.5$, these items would be expected to be classified correctly as "C" items more than 75% of the time. The other values of *a* (0.5 and 1.0) would be expected to yield a "C" classification less than 75% of the time. Thus, values in Table 11 reflect the classification of studied items with $a = 1.5$.

Insert Table 11 about here

For the NO DIF condition, Table 11 reveals that using block level matching results in a large increase over the pooled booklet and complete data analyses in the percentage of items falsely classified as displaying DIF. This is the result of two aspects of the block level matching. Table 6 revealed that, when the studied item did not display DIF, the mean value of $\hat{\Delta}_{MH}$ was substantially more negative (biased) for block level matching than for the pooled booklet approach or for the analysis of the complete data. In addition, Table 8 revealed that $SE(\hat{\Delta}_{MH})$ substantially underestimated the actual variability of $\hat{\Delta}_{MH}$ for the block level approach, especially for shorter blocks. This accounts for the fact that block level matching identified a small percentage of items (2.0% in the NO DIF condition, 0.7% in the DIF condition) as displaying moderate DIF *against the reference group*. Because it does not properly control the number of items falsely identified as displaying DIF (i.e., Type I error rate), the slightly larger number of items correctly identified as "C-" by block level matching does not indicate any superiority or greater sensitivity of the method. It merely reflects that fact that block level matching identifies more items as "C-" whether the item contains true DIF or not.

In general, the results of the pooled booklet approach closely approximate those for the analysis of the complete data. In both the DIF and NO DIF conditions, pooled booklet matching identifies slightly more (< 1%) of the items as "C-" than does the analysis of the complete data. However, given the large decrease in information caused by the BIB design, the behavior of the pooled booklet method is encouraging. It is clearly superior to the other approaches (block level matching, booklet level matching, and the extra-information approach) examined here.

It should be pointed out that the best procedure available in this study for identifying the success of a method of selecting the matching variable for the M-H procedure is the comparison of results for that method with the results for the analysis of the complete data. The results for the analysis of the complete data take into account all that could, in principal, be known about the items at hand; all examinees respond to all items.

Thus, the relatively low power of the analysis of the complete data deserves some further consideration. The accuracy of the complete data results for items not having DIF is very high, but only 40.8 percent of the complete data analyses using the "C-" criterion identified the studied item as having DIF when it did in fact have DIF. The results of Zwick, Thayer, and Wingersky (1993) discussed above suggest that over 75% of these items should have been classified as "C-" items. Two factors interact to cause this result. First, because the items in the simulated tests were generated to have *c*-parameter values other than zero, the assumptions necessary for the M-H statistics are not strictly met. (See Zwick, 1990, for information about the relationship between item response theory and M-H definitions of DIF.) The difference between the method of item generation and the assumptions of the M-H procedures contributes to the poor power of M-H for the analysis of the complete data.

The second factor, discussed above, is that when the items of the test do not follow the Rasch model (as was assumed by Holland & Thayer, 1988) M-H has been found to be sensitive to the difficulty of the studied item (e.g., Donoghue & Allen, 1993; Donoghue, Holland, & Thayer, 1993; Zwick, Thayer, & Wingersky, 1993). Table 5 also demonstrated this sensitivity; as the value of *b* increased, so did the average value of $\hat{\Delta}_{MH}$. Table 12 illustrates that this has a strong, direct effect upon the power of M-H to correctly classify "C-" items.

Insert Table 12 about here

There is relatively little power to detect DIF items with *b*-parameters of 0, and virtually no power for items with $b=1$ or 2. Note that $b=0$ corresponds to a *z*-score of 0.875 for the focal group, and $b=1$ and 2 correspond to *z*-scores of 2.125 and 3.375 respectively. Thus, there is relatively little information about the focal group for these items, and so it is not surprising that M-H has relatively little power to detect DIF. See Zwick, Thayer, and Wingersky (1993) for a discussion of this issue in somewhat more detail. It is not clear at this time whether other DIF methods, such as IRT-based procedures (e.g., Lord, 1980; Raju, 1988; Kim & Cohen, 1991), logistic regression (Swaminathan & Rogers, 1990) or SIBTEST (Shealy & Stout, 1993) can better detect DIF in such difficult items, but it is obviously an sue worthy of further study.

Conclusions

The results of this study give a first indication of the effect of complex item sampling on the M-H procedure. The findings have direct implications for the use of the M-H procedure in NAEP, and give a preliminary indication of what might be expected in future studies examining the analysis of DIF in other settings with complex sampling of items.

In general, these results support the results of Donoghue, Holland, and Thayer (1993), Nelson and Zwick (1989) and Zwick and Grima (1990). Of the models examined by Zwick and Grima, the block level method of matching yields smaller standard errors and more stable estimates than does the booklet method. However, the block analysis results do differ more from the results based on the analysis of the complete data than do the results for the pooled booklet method, a method not considered by Zwick and Grima. This is seen clearly in a comparison of the mean differences in Tables 6 and 10.

It might be expected that results for the extra-information method would most closely reflect those for the analysis of complete data, because the number of levels in the matching variable is larger than for the block, booklet, or pooled booklet matching methods. However, there is some indication that sparseness of data in the large number of 2 X 2 tables used in the extra-information analyses affected the sensitivity of the M-H procedure in this setting. This coincides with the results of Zwick and Ercikan, and can be seen in the mean differences in Table 7 that are larger in magnitude for the extra-information method than for the other methods when the studied item was functioning differentially. From Table 6, the magnitude of $\hat{\Delta}_{MH}$ for the extra-information method is generally smaller than the magnitude for the complete data analysis. Also, Table 9 demonstrates that the $SE(\hat{\Delta}_{MH})$ for the extra-information method tends to underestimate the true variability of $\hat{\Delta}_{MH}$. Thus, the extra-information method cannot be recommended.

The pooled booklet method demonstrated clear advantages over each of the other methods. The booklet method produces three different analyses each with larger standard errors than the pooled booklet method. The extra-information method produces results that differ the most from the results of the complete data analyses. The block method produces results that differ more from the results of the complete data analyses than do the results of the pooled booklet method. In addition, the block method did not control Type I error as well as did the pooled booklet method. Therefore, the pooled booklet approach is recommended for use when items are selected for the examinee according to a BIB design. Some modification of this method might also prove useful for other complex samples of items, although further research is required to verify this.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores, 395-479. Menlo Park, CA: Addison-Wesley Publishing Company.

Beaton, A. E. (1988). Expanding the new design: The NAEP 1985-1986 technical report (No. 17-TR-20). Princeton, NJ: Educational Testing Service.

Bock, R. D., & Mislevy, R. J. (1981). An item response model for matrix-sampling data: The California grade-three assessment. New Directions for Testing and Measurement, 10, 65-90.

Donoghue, 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.

Donoghue, J. R., 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 and H. Wainer (Eds.), Differential item functioning: Theory and practice. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Johnson, E., & Allen, N. L. (1992). The NAEP 1990 technical report (No. 21-TR-20). Princeton, NJ: Educational Testing Service.

Kim, S-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. Applied Psychological Measurement, 15, 269-278.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. Applied Psychological Measurement, 14, 367-386.

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

Nelson, J., & Zwick, R. (1989, April). The Mantel-Haenszel delta difference statistic and its standard error under complex sampling. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

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

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15, 185-197.

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

Zwick, R., & Grima, A. (1990). A proposed policy f.r DIF analysis in NAEP. Personal communication.

Zwick, R., Thayer, D. T., & Wingersky, M. (1993). A simulation study of methods for assessing differential item functioning in computer adaptive tests. Research Report No. RR-93-11. Princeton, NJ: Educational Testing Service.

Table 1*

BIB DESIGN USED IN THE 1990 NAEP MATHEMATICS ASSESSMENT

	Block 1	Block 2	Block 3
Booklet 1	A	B	C
Booklet 2	D	A	E
Booklet 3	F	G	A
Booklet 4	B	E	F
Booklet 5	E	C	G
Booklet 6	G	D	B
Booklet 7	C	F	D

* Other block designations make the structure of the BIB more intuitive. However, the method used above facilitates discussion of the issues in this paper.

Table 2

DATA FOR THE *k*th LEVEL OF THE MANTEL-HAENSZEL MATCHING VARIABLE

Group	Performance on Studied Item		Total
	Passed Item	Failed Item	
Reference	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

Table 3

SCHEMATIC PRESENTATION OF BIB DATA OBTAINED RELEVANT TO BLOCK A

Booklet	Number of Examinees	Block						
		A	B	C	D	E	F	G
		Items A1-A10	Items B1-B10	Items C1-C10	Items D1-D10	Items E1-E10	Items F1-F10	Items G1-G10
1	2000	XXX	XXX	XXX				
2	2000	XXX			XXX	XXX		
3	2000	XXX					XXX	XXX

Quantity	Items in Sum	Number of Possible Score Levels	Symbolic Representation of Number of Score Levels
Block	A1-A10	11 (0, 1, ..., 9, 10)	p_A
Rest of Booklet 1	B1-B10 + C1-C10	21 (0-20)	q_{1A}
Rest of Booklet 2	D1-D10 + E1-E10	21 (0-20)	q_{2A}
Rest of Booklet 3	F1-F10 + G1-G10	21 (0-20)	q_{3A}
Total of Booklet 1	A1-A10 + B1-B10 + C1-C10	31 (0-30)	$k_1 = p_A + q_{1A}$
Total of Booklet 2	A1-A10 + D1-D10 + E1-E10	31 (0-30)	$k_2 = p_A + q_{2A}$
Total of Booklet 3	A1-A10 + F1-F10 + G1-G10	31 (0-30)	$k_3 = p_A + q_{3A}$

Table 4

SCHEMATIC REPRESENTATION OF EXTRA-INFORMATION MATCHING VARIABLE*

		Total Score on Block A			
		0	1	...	10 (p_A)
Total Score on Rest of Booklet 1	0				
	1				
	...				
	20 (q_{1A})				
Total Score on Rest of Booklet 2	0				
	1				
	...				
	20 (q_{2A})				
Total Score on Rest of Booklet 3	0				
	1				
	...				
	20 (q_{3A})				

* Each cell in the table represents a 2X2 table as portrayed in Table 2.

Table 5

MEAN AND (STD. DEV.) OF $\hat{\Delta}_{MH}$ FOR STUDIED ITEM
BASED ON COMPLETE DATA -
NO OTHER DIF ITEMS

NBLK	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
10	-.07 (.37)	-.86 (.86)	750
20	-.04 (.28)	-.88 (.77)	750
30	-.04 (.25)	-.88 (.75)	750

a	b	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
.5	-2	-.02 (.22)	- .73 (.23)	150
.5	-1	.00 (.20)	- .66 (.19)	150
.5	0	.04 (.18)	- .52 (.19)	150
.5	1	.06 (.17)	- .36 (.18)	150
.5	2	.08 (.20)	- .18 (.20)	150
1.0	-2	-.28 (.33)	-1.71 (.28)	150
1.0	-1	-.17 (.24)	-1.42 (.20)	150
1.0	0	-.02 (.19)	- .96 (.18)	150
1.0	1	.07 (.19)	- .43 (.22)	150
1.0	2	.14 (.19)	- .00 (.22)	150
1.5	-2	-.54 (.46)	-2.63 (.38)	150
1.5	-1	-.28 (.23)	-2.07 (.22)	150
1.5	0	-.04 (.21)	-1.17 (.21)	150
1.5	1	.09 (.21)	- .31 (.22)	150
1.5	2	.14 (.24)	.05 (.23)	150

Table 6

MEAN AND (STD. DEV.) OF $\hat{\Delta}_{MH}$ FOR STUDIED ITEMS
BY METHOD OF FORMING MATCHING VARIABLE

	METHOD	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
No Other DIF Items in Block	BLOCK	-.11 (.64)	-.87 (1.13)	2250
	BKLT 1	-.08 (.54)	-.86 (.96)	2250
	BKLT 2	-.07 (.52)	-.90 (.97)	2250
	BKLT 3	-.08 (.53)	-.89 (.98)	2250
	POOLED	-.08 (.42)	-.88 (.92)	2250
	EXTRA-I	-.07 (.32)	-.40 (.68)	2250
	COMPLETE	-.05 (.30)	-.87 (.80)	2250
	METHOD	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
Two Other DIF Items per Block	BLOCK	-.02 (.63)	-.78 (1.11)	2250
	BKLT 1	.03 (.50)	-.78 (.94)	2250
	BKLT 2	.02 (.52)	-.79 (.93)	2250
	BKLT 3	-.00 (.52)	-.79 (.94)	2250
	POOLED	.01 (.40)	-.78 (.88)	2250
	EXTRA-I	.01 (.31)	-.32 (.63)	2250
	COMPLETE	.05 (.28)	-.77 (.75)	2250

Table 7

MEAN AND (STD. DEV.) OF DIFFERENCES BETWEEN $\hat{\Delta}_{MH}$ VALUES
FOR THE METHOD BASED ON COMPLETE DATA AND OTHER METHODS
FOR SELECTED CELLS
($M_{\underline{complete}}$-$M_{\underline{other\ method}}$)

CONDITION/ METHOD			BLOCK		BKLT 1		BKLT 2		BKLT 3		POOLED		EXTRA-I	
a	*b*	NBLK	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF
.5	0	20	-.11 (.08)	-.13 (.09)	-.09 (.29)	-.07 (.24)	.00 (.28)	-.03 (.26)	.00 (.28)	-.00 (.29)	-.03 (.05)	-.04 (.04)	-.02 (.13)	-.32 (.19)
1.0	0	20	.10 (.12)	.04 (.10)	-.00 (.28)	.02 (.24)	-.01 (.28)	.05 (.26)	.09 (.24)	-.03 (.27)	.03 (.06)	.01 (.05)	.10 (.20)	-.72 (.20)
1.5	0	20	.22 (.13)	.11 (.11)	.07 (.27)	-.00 (.29)	.06 (.28)	.07 (.26)	.05 (.27)	.01 (.28)	.06 (.06)	.03 (.05)	.11 (.20)	-.92 (.23)
1.5	-1	20	.65 (.20)	.44 (.17)	.34 (.38)	.17 (.34)	.13 (.33)	.16 (.36)	.22 (.32)	.14 (.37)	.24 (.09)	.16 (.09)	-.09 (.23)	-1.79 (.19)
1.5	1	20	-.21 (.09)	-.26 (.09)	-.09 (.34)	-.17 (.28)	-.02 (.28)	-.09 (.35)	-.09 (.28)	-.00 (.30)	-.07 (.05)	-.10 (.06)	.21 (.25)	-.12 (.20)
1.5	0	10	.31 (.24)	.18 (.20)	.09 (.32)	.06 (.27)	.15 (.30)	.01 (.28)	.08 (.29)	.10 (.27)	.11 (.07)	.05 (.07)	.10 (.14)	.06 (.12)
1.5	0	30	.17 (.10)	.06 (.09)	.08 (.34)	-.04 (.36)	-.01 (.28)	.02 (.26)	.06 (.32)	.05 (.25)	.04 (.05)	.01 (.05)	.16 (.20)	-.96 (.20)

Table 8

MEAN AND (STD. DEV.) OF $SE(\hat{\Delta}_{MH})$ VALUES FOR STUDIED ITEMS
BY METHOD OF FORMING MATCHING VARIABLE

	METHOD	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
No Other DIF Items in Block	BLOCK	.22 (.06)	.21 (.04)	2250
	BKLT 1	.39 (.11)	.37 (.08)	2250
	BKLT 2	.38 (.11)	.37 (.08)	2250
	BKLT 3	.38 (.11)	.37 (.08)	2250
	POOLED	.22 (.06)	.21 (.05)	2250
	EXTRA-I	.14 (.09)	.12 (.08)	2250
	COMPLETE	.22 (.06)	.21 (.05)	2250
	METHOD	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
Two Other DIF Items per Block	BLOCK	.22 (.06)	.21 (.04)	2250
	BKLT 1	.39 (.11)	.37 (.08)	2250
	BKLT 2	.39 (.11)	.37 (.08)	2250
	BKLT 3	.39 (.11)	.37 (.08)	2250
	POOLED	.22 (.06)	.22 (.05)	2250
	EXTRA-I	.15 (.09)	.12 (.08)	2250
	COMPLETE	.22 (.06)	.22 (.05)	2250

Table 9

MEAN RATIOS OF OBSERVED SD($\hat{\Delta}_{MH}$) VALUES TO THE AVERAGE $SE(\hat{\Delta}_{MH})$ FOR STUDIED ITEMS BY METHOD OF FORMING THE MATCHING VARIABLE AND NUMBER OF ITEMS IN THE BLOCK (SD/AVERAGE *SE*)

CONDITION/NBLK	10		20		30	
METHOD	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF
BLOCK	1.25 (.21)	1.22 (.11)	1.10 (.14)	1.10 (.12)	1.00 (.10)	1.06 (.11)
BKLT 1	1.03 (.10)	1.02 (.10)	.98 (.11)	1.00 (.05)	1.02 (.11)	.96 (.09)
BKLT 2	.98 (.12)	1.00 (.12)	1.02 (.08)	1.02 (.08)	.97 (.09)	1.04 (.07)
BKLT 3	1.02 (.09)	.99 (.10)	.99 (.08)	1.01 (.12)	.98 (.13)	1.00 (.10)
POOLED	.98 (.14)	1.01 (.08)	1.03 (.08)	1.03 (.10)	.97 (.12)	1.04 (.12)
EXTRA-I	1.03 (.11)	1.05 (.09)	1.50 (.36)	1.69 (.35)	1.26 (.20)	1.33 (.16)
COMPLETE	.97 (.14)	1.00 (.06)	1.02 (.09)	1.01 (.09)	.97 (.11)	1.04 (.13)

Table 10

MEAN AND (STD. DEV.) OF DIFFERENCES BETWEEN $\hat{\Delta}_{MH}$ VALUES
FOR THE METHOD BASED ON COMPLETE DATA
AND BLOCK AND POOLED BOOKLET ANALYSES
NO OTHER DIF ITEMS
($M_{\underline{complete}}$-$M_{\underline{other\ method}}$)

	NBLK	NO DIF	DIF	# OF REPLICATIONS FOR EACH MEAN
Block Level Analysis	10	.03 (.48)	-.02 (.42)	750
	20	.07 (.42)	.00 (.36)	750
	30	.08 (.37)	.02 (.31)	750
Pooled Booklet Analysis	10	.02 (.20)	.00 (.17)	750
	20	.03 (.16)	.01 (.14)	750
	30	.03 (.14)	.01 (.12)	750

Table 11

PERCENT OF ITEMS CLASSIFIED INTO ETS DIF CATEGORIES*
NO OTHER DIF ITEMS--*a*=1.5
(OUT OF 750 ANALYSES)

DIF IN ITEM	METHOD	C-	B-	A- & A+	B+	C+
NO DIF	BLOCK	6.4	17.7	73.9	2.0	0.0
	POOLED	1.1	7.6	91.2	0.1	0.0
	COMPLETE	0.3	2.7	97.1	0.0	0.0
DIF	BLOCK	43.5	14.9	40.9	0.7	0.0
	POOLED	41.3	15.9	42.8	0.0	0.0
	COMPLETE	40.8	15.2	44.0	0.0	0.0

* "C" indicates substantial evidence of DIF, "B" indicates moderate evidence of DIF, and "A" indicates little or no evidence of DIF. Minus suffix indicates $\hat{\Delta}_{MH} < 0$ (poorer performance by the focal group, conditional on matching variable). Plus indicates $\hat{\Delta}_{MH} > 0$ (poorer performance by the reference group, conditional on matching variable). See text for full category definitions.

Table 12

PERCENT OF ITEMS CLASSIFIED INTO ETS DIF CATEGORIES*
BY DIFFICULTY OF STUDIED ITEM--ANALYSIS OF COMPLETE DATA,
NO OTHER DIF ITEMS, a=1.5
(OUT OF 150 ANALYSES)

DIF IN ITEM	b	C-	B-	A- & A+	B+	C+
NO DIF	-2	1.3	13.3	85.3	0.0	0.0
	-1	0.0	0.0	100.0	0.0	0.0
	0	0.0	0.0	100.0	0.0	0.0
	1	0.0	0.0	100.0	0.0	0.0
	2	0.0	0.0	100.0	0.0	0.0
DIF	-2	98.7	1.3	0.0	0.0	0.0
	-1	99.3	0.7	0.0	0.0	0.0
	0	6.0	73.3	20.7	0.0	0.0
	1	0.0	0.7	99.3	0.0	0.0
	2	0.0	0.0	100.0	0.0	0.0

* "C" indicates substantial evidence of DIF, "B" indicates moderate evidence of DIF, and "A" indicates little or no evidence of DIF. Minus suffix indicates $\hat{\Delta}_{MH} < 0$ (poorer performance by the focal group, conditional on matching variable). Plus indicates $\hat{\Delta}_{MH} > 0$ (poorer performance by the reference group, conditional on matching variable). See text for full category definitions.

Table A1

MEAN AND (STD. DEV.) OF DIFFERENCES BETWEEN $\hat{\Delta}_{MH}$ VALUES FOR THE METHOD BASED ON COMPLETE DATA AND ON BLOCK ANALYSES-NO OTHER DIF ITEMS

($M_{\underline{complete}}$-$M_{\underline{block}}$)

CONDITION/NBLK		10		20		30	
a	*b*	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF
.5	-2	-.06 (.17)	-.05 (.19)	-.00 (.11)	-.01 (.13)	.00 (.09)	.01 (.10)
.5	-1	-.10 (.17)	-.12 (.15)	-.06 (.10)	-.07 (.10)	-.04 (.09)	-.04 (.06)
.5	0	-.17 (.15)	-.18 (.15)	-.11 (.08)	-.13 (.09)	-.08 (.07)	-.10 (.06)
.5	1	-.27 (.12)	-.29 (.14)	-.20 (.09)	-.22 (.08)	-.15 (.05)	-.17 (.05)
.5	2	-.38 (.13)	-.40 (.10)	-.29 (.08)	-.29 (.07)	-.23 (.05)	-.24 (.05)
1.0	-2	.52 (.27)	.44 (.24)	.53 (.21)	.43 (.21)	.49 (.17)	.40 (.13)
1.0	-1	.37 (.24)	.29 (.21)	.33 (.16)	.25 (.14)	.31 (.11)	.23 (.10)
1.0	0	.12 (.18)	.08 (.18)	.10 (.12)	.04 (.10)	.08 (.08)	.04 (.07)
1.0	1	-.22 (.13)	-.25 (.15)	-.18 (.08)	-.20 (.08)	-.15 (.06)	-.17 (.07)
1.0	2	-.48 (.12)	-.49 (.13)	-.38 (.07)	-.40 (.07)	-.28 (.05)	-.31 (.06)
1.5	-2	.97 (.36)	.83 (.28)	.99 (.32)	.79 (.29)	.95 (.31)	.72 (.25)
1.5	-1	.71 (.27)	.52 (.22)	.65 (.20)	.44 (.17)	.60 (.15)	.38 (.12)
1.5	0	.31 (.24)	.18 (.20)	.22 (.13)	.11 (.11)	.17 (.10)	.06 (.09)
1.5	1	-.25 (.14)	-.30 (.14)	-.21 (.09)	-.26 (.09)	-.19 (.07)	-.23 (.06)
1.5	2	-.57 (.09)	-.56 (.10)	-.42 (.08)	-.43 (.08)	-.34 (.06)	-.33 (.06)

Table A2

MEAN AND (STD. DEV.) OF DIFFERENCES BETWEEN $\hat{\Delta}_{MH}$ VALUES FOR THE METHOD BASED ON COMPLETE DATA AND ON POOLED BOOKLET ANALYSES-NO OTHER DIF ITEMS

($M_{\underline{complete}}$-$M_{\underline{block}}$)

CONDITION/NBLK		10		20		30	
a	b	NO DIF	DIF	NO DIF	DIF	NO DIF	DIF
.5	-2	-.01 (.08)	-.00 (.06)	-.01 (.06)	-.00 (.07)	-.00 (.06)	.00 (.07)
.5	-1	-.04 (.05)	-.04 (.05)	-.01 (.05)	-.02 (.05)	-.01 (.05)	-.01 (.05)
.5	0	-.06 (.06)	-.06 (.04)	-.03 (.05)	-.04 (.04)	-.03 (.05)	-.03 (.05)
.5	1	-.09 (.04)	-.11 (.04)	-.06 (.05)	-.08 (.05)	-.04 (.04)	-.05 (.04)
.5	2	-.14 (.05)	-.14 (.04)	-.10 (.04)	-.10 (.05)	-.07 (.05)	-.07 (.05)
1.0	-2	.25 (.11)	.20 (.11)	.20 (.12)	.17 (.10)	.17 (.11)	.14 (.08)
1.0	-1	.16 (.10)	.12 (.07)	.13 (.08)	.09 (.06)	.10 (.06)	.08 (.06)
1.0	0	.04 (.05)	.03 (.06)	.03 (.06)	.01 (.05)	.02 (.05)	.01 (.04)
1.0	1	-.09 (.05)	-.10 (.05)	-.06 (.05)	-.07 (.05)	-.04 (.05)	-.05 (.06)
1.0	2	-.18 (.04)	-.17 (.05)	-.13 (.05)	-.13 (.05)	-.08 (.05)	-.09 (.04)
1.5	-2	.43 (.18)	.37 (.15)	.40 (.16)	.30 (.16)	.33 (.19)	.25 (.18)
1.5	-1	.30 (.10)	.22 (.10)	.24 (.09)	.16 (.09)	.20 (.07)	.12 (.06)
1.5	0	.11 (.07)	.05 (.07)	.06 (.06)	.03 (.05)	.04 (.05)	.01 (.05)
1.5	1	-.11 (.05)	-.13 (.06)	-.07 (.05)	-.10 (.06)	-.07 (.05)	-.07 (.05)
1.5	2	-.20 (.05)	-.20 (.05)	-.13 (.05)	-.14 (.05)	-.10 (.05)	-.09 (.04)