DOCUMENT RESUME

ED 387 526 TM 023 986

AUTHOR	Dorans, Neil J.; Holland, Paul W.
TITLE	DIF Detection and Description: Mantel-Haenszel and Standardization.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-92-10
PUB DATE	Jan 92
NOTE	43p.; Version of a paper presented at the Educational Testing Service/AFHRL Conference (Princeton, NJ, October 6, 1989).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Comparative Analysis; *Distractors (Tests); *Identification; *Item Bias; Item Response Theory; Mathematical Models; *Test Items
IDENTIFIERS	Contingency Tables; *Mantel Haenszel Procedure; Rasch Model; *Standardization

ABSTRACT

At the Educational Testing Service, the Mantel-Haenszel procedure is used for differential item functioning (DIF) detection, and the standardization procedure is used to describe DIF. This report describes these procedures. First, an important distinction is made between DIF and impact, pointing to the need to compare the comparable.. Then, these two contingency table DIF procedures are described in some detail, first in terms of their own origins as DIF procedures, and then from a common framework that points out similarities and differences. The relationship between the Mantel-Haenszel procedure and item response theory models, in general, and the Rasch model, in particular, is discussed. The utility of the standardization approach for assessing differential distractor functioning is described. Several issues in applied DIF analyses are discussed, including inclusion of the studied item in the matchinr variable and refinement of the matching variable. Future research topics dealing with the matching variable, the studied variable, and the group variable are also discussed. (Contains 5 figures and 48 references.) (SLD)

RR-92-10

RESEARCH REPORT

DIF DETECTION AND DESCRIPTION: MANTEL-HAENSZEL AND STANDARDIZATION

Neil J. Dorans
Paul W. Holland

Educational Testing Service
Princeton, New Jersey
January 1992

DIF Detection and Description: Mantel-Haenszel and Standardization[1,2]

Neil J. Dorans and Paul W. Holland

Educational Testing Service

November 1991

[1] An earlier version of this report was presented at ETS/AFHRL Conference, *Differential Item Functioning:Theory and Practice,* Educational Testing Service, Princeton, NJ, October 6, 1989. This report will appear as Chapter 3 in P. W. Holland & H. Wainer .(Eds.), *Differential item functioning,* which will be published by Lawrence Erlbaum Associates in 1992. Citations for this work should be to the chapter version.

[2] The opinions expressed in this report are the those of one or both of the authors and should not be misconstrued to represent official policy of the Educational Testing Service. The authors are grateful to Alicia Schmitt and Howard Wainer for their careful reviews of earlier versions of this report.

Abstract

At the Educational Testing Service, the Mantel-Haenszel procedure is used for differential item functioning (DIF) detection and the standardization procedure is used to describe DIF. This report describes these procedures. First, an important distinction is made between DIF and Impact, pointing the need to compare the comparable. Then, these two contingency table DIF procedures are described in some detail, first in terms of their own origins as DIF procedures, and then from a common framework that points out similarities and differences. The relationship between the Mantel-Haenszel procedure and IRT models in general and the Rasch model, in particular, is discussed. The utility of the standardization approach for assessing differential distractor functioning is described. Several issues in applied DIF analyses are discussed including inclusion of the studied item in the matching variable, and refinement of the matching variable. Future research topics dealing with the matching variable, the studied variable and the group variable are also discussed.

DIF Detection and Description: Mantel-Haenszel and Standaıdization

Neil J. Dorans and Paul W. Holland

Differential item functioning (DIF) refers to a psychometric difference in how an item functions for two groups. DIF refers to a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning from difference between groups.

In the first chapter of the book, *Handbook of Methods for Detecting Test Bias*, Shepard (1982) defines what was then called item bias and is now referred to as DIF as psychometric features of the item that can misrepresent the competence of one group. She provides an understanding of the meaning of DIF by presenting some conceptual definitions of the term, including:

> *An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered.* (Scheuneman, 1975, p. 2)

This definition by Scheuneman may be the earliest contingency table definition of DIF. It is the definition underlying the observed score DIF approaches described in this report.

Lord (1980) provides the item response theory definition of DIF:

> *If each test item in a test had exactly the same item response function in every group, then people of the same ability or skill would have exactly the same chance of getting the item right, regardless of their group membership. Such a test would be completely unbiased. If on the other hand, an item has a different item response function for one group than for another, it is clear that the item is biased. (p. 212)*

This item response theory definition underlies the DIF procedures described in Thissen, Steinberg and Wainer (in press).

Thissen (1987), in his discussion of a series of DIF papers dealing with DIF on the Scholastic Aptitude Test (SAT) that are contained in Schmitt and Dorans (1987), adds to these definitions by referring to DIF as:

> *...an expression which describes a serious threat to the validity of tests used to measure the aptitude of members of different populations or groups. Some test items may simply perform differently for examinees drawn from one group or another or they may measure "different things" for members of one group as opposed to members of another. Tests comparing such items may have reduced validity for between-group comparison, because their scores may be indicative of a variety of attributes other than those the test is intended to measure.* (p. 1)

Statistical methods used to identify DIF are defined by Shepard (1982) as: " *internal methods designed to ensure that the meaning, which individual items attribute to the total test, is the same for all subgroups.* (p. 23). A variety of methods have been used since the 1950s.. Two methods presently employed at the Educational Testing Service for DIF assessment are the standardization approach (Dorans & Kulick, 1986) and the Mantel-Haenszel approach (Holland & Thayer, 1988). Both procedures compare matched or comparable groups. This report describes these two procedures in some detail.

The structure of the report is as follows. DIF is contrasted with impact via Simpson's paradox, which demonstrates the importance of matching in DIF studies. Then a definition of DIF is offered. The Mantel-Haenszel(MH) procedure is described as a statistically powerful method for detecting DIF, and the standardization approach is described as a flexible procedure for describing DIF. A common framework from which to view these two related procedures is then presented. Then, the relationship between the MH procedure and the Rasch model under the condition that the Rasch model is appropriate for the data is discussed. Next, the utility of the standardization approach for assessing differential distractor functioning is described. Some issues in applied DIF analyses are discussed. Finally, future directions in DIF analyses are considered.

1. DIF Not Impact

It is important to make a distinction between DIF and impact. Impact refers to a difference in performance between two intact groups. Impact is everywhere in test and item data because individuals differ with respect to the developed abilities measured by items and tests, and intact groups, such as those defined by ethnicity and gender differ with respect to the distributions of developed ability among their members. For example, on a typical SAT-Mathematics item it is usually the case that Asian-Americans, as a group, score higher than Whites, males score higher than females, and juniors and seniors score higher than junior high school students. This difference in performance is called impact. Frequently, impact on any given item is consistent with impact on other items of the same type. In fact, impact at the item level is frequently explained by impact across all items of similar type or impact at the total score level.

In contrast to impact, which can often be explained by stable consistent differences in examinee ability distributions across groups, DIF refers to differences in item functioning after groups have been matched with respect to the ability or attribute that the item purportedly measures. Unlike impact, where differences in item performance reflect differences in overall ability distributions, DIF is an unexpected difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and test on which it appears.

1.1 Simpson's Paradox

Simpson's paradox (Simpson, 1951) illustrates why we should compare the comparable, as is done in DIF analyses. The following table summarizes the performance of two hypothetical groups, A and B, on an imaginary item.

Group A			Group B		
N_m	N_{cm}	N_{cm}/N_m	N_m	N_{cm}	N_{cm}/N_m
400	40	.10	1000	200	.20
1000	500	.50	1000	600	.60
1000	900	.90	400	400	1.0
2400	1440	.60	2400	1200	.50

This table contains four rows and six columns of numbers The first three columns pertain to group A, while the last three pertain to group B. The first three rows pertain to three different ability levels ranging from the lowest to the highest, while the fourth row sums across ability levels. (In the case of the the third and sixth columns, the sum in the fourth row is a weighted sum.) The symbols N_m, N_{cm}, and N_{cm}/N_m refer to the number of people at the ability level **m**, the number of people at ability level **m** who answered the item correctly, and the proportion at ability level **m** who answered the item correctly, respectively.

Of the 2,400 examinees in group A, 1,440 or 60% answered the item correctly. In contrast, only 50%, 1,200 of 2,400, of group B answered the item correctly. The impact on this item is .6 - .5 = .1 in favor of group A.

Upon closer examination, however, the ratio N_{cm}/N_m at each of the three ability levels for group A is actually .1 lower than the corresponding ratio for group B. These conditional proportions are .1, .5, and .9 for group A, and .2, .6. and 1.0 for group B. Hence, when we compare the comparable at each ability level **m**, we find that this item actually favors group B over group A, not vice versa as suggested by impact. This contradiction between impact and DIF is due to unequal distributions of ability in groups A and B, as seen in the N_m columns. This imaginary item actually disadvantages group A, but since group A is more able than group B, the overall impact suggests that the item favors group A.

Simpson's paradox has a rich history in the statistical literature (e.g., Blyth, 1972; Wagner, 1982; Yule, 1903). Recently, Wainer (1986) illustrated how this paradox affects the interpretation of changes in SAT mean scores over time. Simpson's paradox illustrates the importance of comparing the comparable. Both the standardization approach (Dorans & Kulick, 1983, 1986), which has been used on the SAT since 1982, and the Mantel-Haenszel method (Holland & Thayer, 1988), which has been used with most ETS testing programs since 1987, emphasize the importance of comparing the comparable. In practice, both approaches use equal ability as measured by total test score as a measure of comparability. They share a common definition of Null DIF, namely that there is no differential item functioning between groups after they have been matched on total score. In theory, both procedures are flexible enough to match on more than total score (see last portion of this report for a discussion of this issue). In practice, matching is typically based on a single total score.

These two DIF assessment procedures are highly related and complement each other well. The Mantel-Haenszel is a statistically powerful technique for detecting DIF. Standardization is a very flexible, easily understood descriptive procedure that is particularly suited for assessing plausible and implausible explanations of DIF.

2. Mantel-Haenszel: Testing the Constant Odds Ratio Hypothesis Version of DIF

In their seminal paper, Mantel and Haenszel (1959) introduced a new procedure for the study of matched groups. Holland (1985) and later Holland and Thayer (1988) adapted the procedure for use in assessing differential item functioning. This adaptation is used at the Educational Testing Service as the primary DIF detection device. The basic data used by the MH method are in the form of M 2 X 2 contingency tables or one large three dimensional **2-by-2-M** table.

2.1 The 2-by-2-by-M Contingency Table

Under rights scoring for the items in which responses are coded as either correct or incorrect (including omissions), counts of rights and wrongs on each item can be arranged into a **2-by-2-by-M** contingency table for each item being studied. There are two levels for group; the focal group that is the focus of analysis and the reference group that serves as a basis for comparison for the focal group. At ETS, the current practice is to do analyses in which Whites are the reference group, and Blacks, Hispanics, Asian-Americans, and Native Americans, serve as the focal groups, and analyses in which females are the focal group and males are the reference group. There are two levels for item response; right or wrong, and there are M score levels on the matching variable, e.g. total score. Finally, the item being analyzed is referred to as the studied item. The **2(groups)-by-2(item scores)-by-M(score levels)** contingency table for each item can be viewed in *2-by-2 slices* (there are M slices per item) as shown below:

	Item Score		
Group	Right	Wrong	Total
Focal Group (f)	R_{fm}	W_{fm}	N_{fm}
Reference Group (r)	R_{rm}	W_{rm}	N_{rm}
Total Group (t)	R_{tm}	W_{tm}	N_{tm}

The null DIF hypothesis[1] for the Mantel-Haenszel method can be expressed as

$$H_0: [R_{rm}/W_{rm}] / [R_{fm}/W_{fm}] = 1 \qquad m = 1,\ldots, M,$$

or alternatively,

$$H_0: [R_{rm}/W_{rm}] = [R_{fm}/W_{fm}] \qquad m = 1,\ldots,M.$$

In other words, the odds of getting the item correct at a given level of the matching variable is the same in both the focal group and the reference group, across all M levels of the matching variable.

[1] Note that in stating hypotheses we have not distinguished between population and sample quantities. All of our hypotheses should read as relations among the expectations of the indicated statistics.

2.2 The Constant Odds Ratio Hypothesis

In their original work, Mantel and Haenszel (1959) developed a chi-square test of the null DIF hypothesis against a particular alternative hypothesis known as the constant odds ratio hypothesis,

$$H_a: [R_{rm}/W_{rm}] = \alpha\,[R_{fm}/W_{fm}] \qquad m = 1,\ldots,M \text{ and } \alpha \neq 1.$$

Note that when $\alpha = 1$, the alternative hypothesis reduces to the null DIF hypothesis. The parameter α is called the *common odds ratio* in the **M** 2-by-2 tables because under H_a, the value of α is the odds ratio that is the same for all **m**,

$$\alpha_m = [R_{rm}/W_{rm}]/[R_{fm}/W_{fm}] = [R_{rm}W_{fm}]/[R_{fm}W_{rm}].$$

2.3 Chi Square Test Statistic

There is a chi-square test associated with the MH approach, namely a test of the null hypothesis, H_0: $\alpha_m = 1$,

$$MH\text{-}\chi^2 = [\,|\Sigma_m R_{rm} - \Sigma_m E(R_{rm})| - .5]^2/\Sigma_m Var(R_{rm}),$$

where,

$$E(R_{rm}) = E(R_{rm} \mid \alpha = 1) = N_{rm}R_{tm}/N_{tm},$$

$$Var(R_{rm}) = Var(R_{rm} \mid \alpha = 1) = [N_{rm}R_{tm}N_{fm}W_{tm}]/[N_{tm}^2(N_{tm} - 1)],$$

and where the -.5 in the expression for $MH\text{-}\chi^2$ serves as a continuity correction to improve the accuracy of the chi-square percentage points as approximations to the observed significance levels. The quantity $MH\text{-}\chi^2$ is approximately distributed as a chi-square with one degree of freedom.

Holland and Thayer (1988) report:

> *...that a test based on $MH\text{-}\chi^2$ is the uniformly most powerful unbiased test of H_o versus H_a. Hence no other test can have higher power somewhere in H_a than the one based on $MH\text{-}\chi^2$ unless the other test violates the size constraint on the null hypothesis or has lower power than the test's size somewhere else on H_a.* (p. 134)

In other words, the MH approach is the statistical test possessing the most statistical power for detecting departures from the null DIF hypothesis that are consistent with the constant odds ratio hypothesis.

2.4 Estimate of Constant Odds Ratio

Mantel and Haenszel also provided an estimate of the constant odds-ratio,

$$\alpha_{MH} = [\Sigma_m R_{rm} W_{fm}/N_{tm}] / [\Sigma_m R_{fm} W_{rm}/N_{tm}] \,.$$

This estimate is an estimate of DIF effect size in a metric that ranges from 0 to ∞ with a value of 1 indicating null DIF. This odds-ratio metric is not particularly meaningful to test developers who are used to working with numbers in an item difficulty scale. In general, odds are converted to log odds because the latter are symmetric around zero and easier to interpret.

2.5 MH DIF in Item Difficulty Metrics

At ETS, test developers are used to working with item difficulty estimates in the "delta metric", which has a mean of 13 and a standard deviation of 4. To obtain a delta, the proportion correct (p) is converted to a z-score via a p-to-z transformation using the inverse of the normal cumulative function, followed by a linear transformation to a metric with a mean of 13 and a standard deviation of 4 via:

$$\Delta = 13 - 4\{\Phi^{-1}(p)\},$$

such that large values of Δ correspond to difficult items, while easy items have small values of delta. Holland and Thayer (1985) converted α_{MH} into a difference in deltas via:

$$\text{MH D-DIF} = -2.35 \ln[\alpha_{MH}].$$

Note that positive values of MH D-DIF favor the focal group, while negative values favor the reference group.

Another metric that is used more universally to describe item difficulty is the p-metric, percent correct or proportion correct metric. The α_{MH} can also be expressed in this metric,

$$\text{MH P-DIF} = P_f - P_r{\dagger},$$

where,

$$P_r\dagger = [\alpha_{MH}P_f] / [(1-P_f) + \alpha_{MH}P_f] ,$$

which can be thought of as a predicted proportions correct in the reference group based on the MH odds-ratio, and P_f is the proportion correct observed in the focal group.

2.6 Standard Error of the Mantel-Haenszel DIF Indices

A useful, approximate standard error for the log of the Mantel-Haenszel odds-ratio estimator was developed by Robins, Breslow and Greenland (1986) and, in the equivalent form used here, by Phillips and Holland (1987). This expression may be multiplied by 2.35 to yield an estimated standard error for MH D-DIF,

$$\text{SE (MH D-DIF)} = \{2.35/C\}^*\{\Sigma_m[(R_{rm}W_{fm} + \alpha_{MH}W_{rm}R_{fm}) * [R_{rm} + W_{fm} + \alpha_{MH}(W_{rm} + R_{fm})]/(2N_{tm}^2)]\}^{.5} ,$$

where,

$$C = \Sigma_m R_{rm}W_{fm}/N_{tm} \cdot$$

The standard error for MH P-DIF, derived in Holland(1989), is

$$\text{SE (MH P-DIF)} = \{(1-K)^2P_f(1-P_f)/N_f + 2K(1-K)P_f(1-P_f)/N_f + K^2[P_f(1-P_f))]^2 [\text{SE(MH D-DIF)}/(2.35)]^2\}^{.5} ,$$

where,

$$K = \alpha_{MH}/(1 - P_f + \alpha_{MH}P_f)^2 ,$$

and N_f is the total number of examinees in the focal group.

2.7 ETS DIF Classification Rules

To use the MH D-DIF measure to identify test items that exhibit varying degrees of DIF, a classification scheme was developed at ETS for use in test development that puts items into one of three categories -- negligible DIF (A), intermediate DIF (B), and large DIF (C). Items are classified as A for a particular combination of reference and focal groups if either MH D-DIF is not statistically different from zero or if the magnitude of the MH D-DIF values is less than one delta unit in absolute value. Items are classified as C if MH D-

DIF both exceeds 1.5 in absolute value and is statistically significantly larger than 1.0 in absolute value. All other items are classified as category B. In both categories A and C statistical significance is at the 5% level for a single item. Presently an item can have up to five different MH D-DIF values associated with it, one for each of five possible combinations of focal and reference groups. An item is currently assigned the lowest letter grade from all the DIF analyses performed on it.

2.8 The MH Procedure and The Rasch Model

Holland and Thayer (1988) point out a close connection between "chi-square" types of DIF procedures, such as the MH procedure and "theoretically preferred" methods based on item response theory models, such as those described by Thissen, Steinberg and Wainer (in press). They draw this close connection in fairly abstract terms using a very general class of item response theory models. The interested reader should consult the original source for the mathematical details. To make matters concrete Holland and Thayer show how the Rasch model and the MH procedure are related when the assumptions underlying the Rasch model fit the data. In particular, they demonstrate that under the Rasch model the constant odds ratio hypothesis holds exactly in the population if: (1) all items in the matching criterion, with the possible exception of the studied item, are free of DIF; (2) the criterion for matching is a number-right score that includes the studied item; and (3) the data are random samples from the reference and focal populations. It is only under these special conditions, some of which are strong, particularly the assumption that the Rasch model fits the data, that MH and Rasch model have a special relationship. It is important to realize that the Holland and Thayer analysis does not imply that the Rasch model and MH procedure are always intimately related. Instead, Holland and Thayer (1988) used the MH procedure and the Rasch model to relate the chi-square procedures and the the item response theory procedures under special conditions. In the process, they determined the need to include the studied item in the matching criterion, which has implications for DIF applications and future research, both of which will be discussed later.

3. Standardization: A Flexible Method for Describing DIF

In the early eighties, Dorans (1982) reviewed a number of item bias studies that had been conducted on SAT data in the late seventies. These studies had used the Angoff and Ford (1973) delta-plot methodology and, in some cases, a log-linear method. The delta-plot method can be justified from a one-parameter normal ogive item response theory model, and as such, is of as limited applicability to multiple-choice item data as the Rasch model. DIF detection with either the Rasch model or the delta-plot model is confounded with lack of model fit, a confounding that occurs frequently because items do not have a common discrimination parameter. The log-linear approach employed in those early SAT studies was flawed because the conditioning variable was too coarsely grouped, a practice we refer to as *fat matching*. Taken to its extreme, fat matching leads to a single level for the matching variable, which converts DIF studies into impact studies. Dorans (1982) concluded that a new method was needed.

Large data sets are often associated with SAT test forms. Given large SAT data sets and a desire to avoid contamination caused by model misfit, Dorans and Kulick (1983) decided to not employ IRT models. Instead, they opted for an IRT-like approach that compared empirical item response curves in which a total score was used as an estimate of ability. Summarizing these numerous non-parametric item test regressions via some numerical index seemed to be essential if this procedure was to become practical. They were steered in the direction of standardization via the Alderman and Holland (1981) report on DIF assessment for the Test of English as a Foreign Language (TOEFL).

According to the standardization method, an item is exhibiting DIF when the expected performance on an item differs for examinees of equal ability from different groups. Expected performance on an item can be operationalized by non-parametric item test regressions. Differences in empirical item test regressions are indicative of DIF.

One of the main principles underlying the standardization approach to DIF assessment is to use all available appropriate data to estimate the conditional item performance of each group at each level of the matching variable. The matching done by standardization and Mantel-Haenszel does not require the use of stratified sampling procedures that yield equal numbers of examinees at a given score level across groups. In fact, throwing away data in this fashion just leads to poorer estimates of effect sizes that have larger standard errors associated with them than effect sizes based on all the data.

The first step in the standardization analysis is to use all available data to estimate non-parametric item test regressions in the reference group and in the focal group. Let $E_f(I \mid M)$ define the empirical item test regression for the focal group *f*, and let $E_r(I \mid M)$ define the empirical item test regression for the reference group *r*, where I is the item score variable and M is the matching variable. The definition of DIF employed by the standardization approach implies that $E_f(I \mid M) = E_r(I \mid M)$.

The most detailed definition of DIF is at the individual score level, *m*,

$$D_m = E_{fm} - E_{rm}$$

where, E_{fm} and E_{rm} are realizations of the item-test regressions at score level m. The D_m are the fundamental measures of DIF according to the standardization method because these quantities are differences in item performance between focal group and reference group members who are matched with respect to the attribute measured by the test. Any differences that exist after matching cannot be explained or accounted for by ability differences. These are unexpected differences as opposed to those expected given ability differences. Plots of these differences, as well as plots of $E_f(I \mid M)$ and $E_r(I \mid M)$, provide visual descriptions of DIF in fine detail. Figures 1 and 2 are sample plots of non-parametric item test regressions and differences for an actual SAT item, which exhibits considerable DIF. In contrast, Figures 3 and 4 are item test regressions for an actual SAT item which exhibits minimal DIF.

Visual analysis is an important component of the standardization approach. Figure 1 comes from the first study to use standardization to do DIF analyses on the SAT (Dorans & Kulick, 1983). In that study, there were 21,209 female examinees in the focal group and 21,285 male examinees in the reference group. In Figure 1, E_{fm} and E_{rm} are presented in a percent correct metric, ranging from 0 to 100, while the matching variable is score on the familiar 200-to-800 College Board scale. Each point in the plot represents the conditional item mean score (under rights scoring) at each scaled score level. This plot and the corresponding difference plot in Figure 2 provide detailed visual descriptions of difference and similarities of focal and reference group performance on the item at each of the 61 scale score levels ranging from 200 to 800 in 10 point increments.

The content for this item, which appeared on the December 1977 form of the SAT, reveals why there is such large DIF on this item. It is a verbal analogy item,

DECOY : DUCK :: (A) net : butterfly (B) web : spider
(C) lure : fish (D) lasso : rope (E) detour : shortcut.

This edition of the SAT was assembled prior to the institution of the ETS Test Sensitivity Guidelines (see Ramsey, in press), which screen items for content or language that is offensive or could be detrimental to the performance of ethnic or gender subgroups. Had such guidelines been in place, this item may never have appeared in a final edition of the SAT because a casual examination of the item reveals that knowledge of hunting and fishing jargon probably influence performance on this item. Sex differences with respect to familiarity with this jargon probably accounts for why males outperform matched females at difference of 15% to 20% at each SAT-Verbal scaled score level between 250 and 500. For example at a scaled score level of 300, over 60% of the males answer the item correctly, while only 40% of the females choose the correct response option. Clearly, this is a very easy item for males that is somewhat harder for females, an item that exhibits substantial DIF, and a high DIF item that is biased against females.

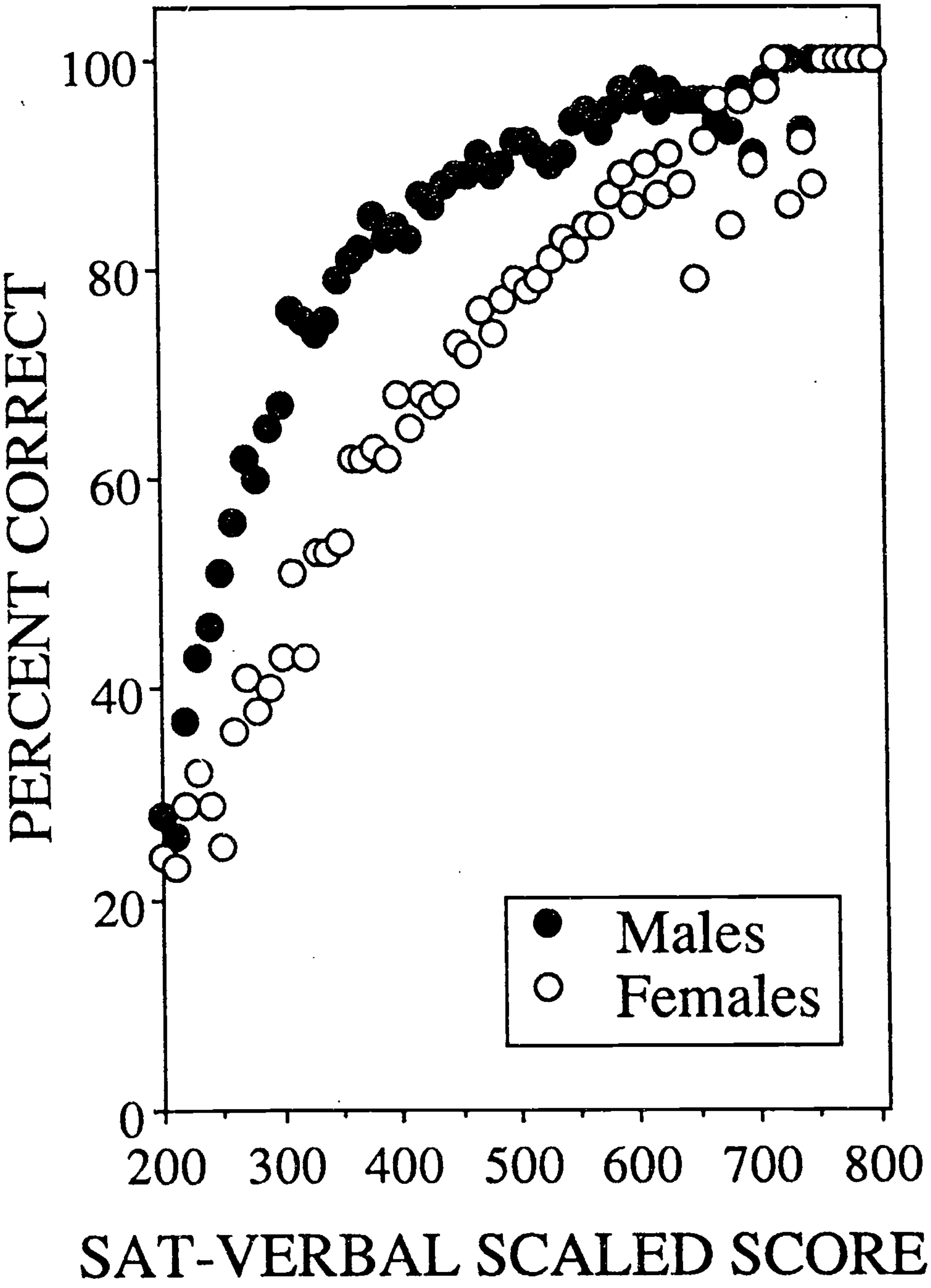
PERCENT CORRECT
0
20
40
60
80
100
200
300
400
500
600
700
800
SAT-VERBAL SCALED SCORE
● Males
○ Females

Figure 1

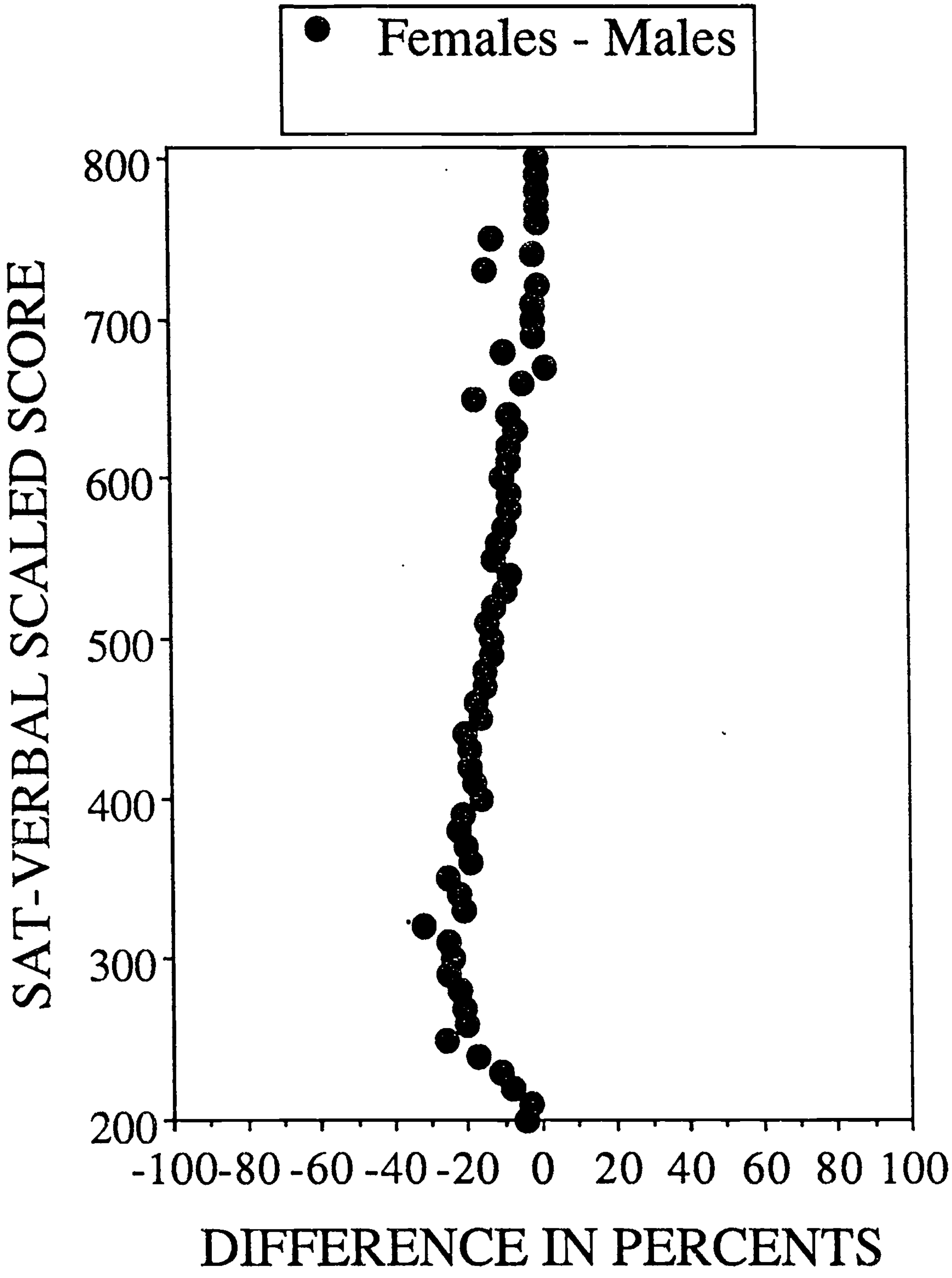
● Females - Males
SAT-VERBAL SCALED SCORE
800
700
600
500
400
300
200
-100
-80
-60
-40
-20
0
20
40
60
80
100
DIFFERENCE IN PERCENTS

Figure 2

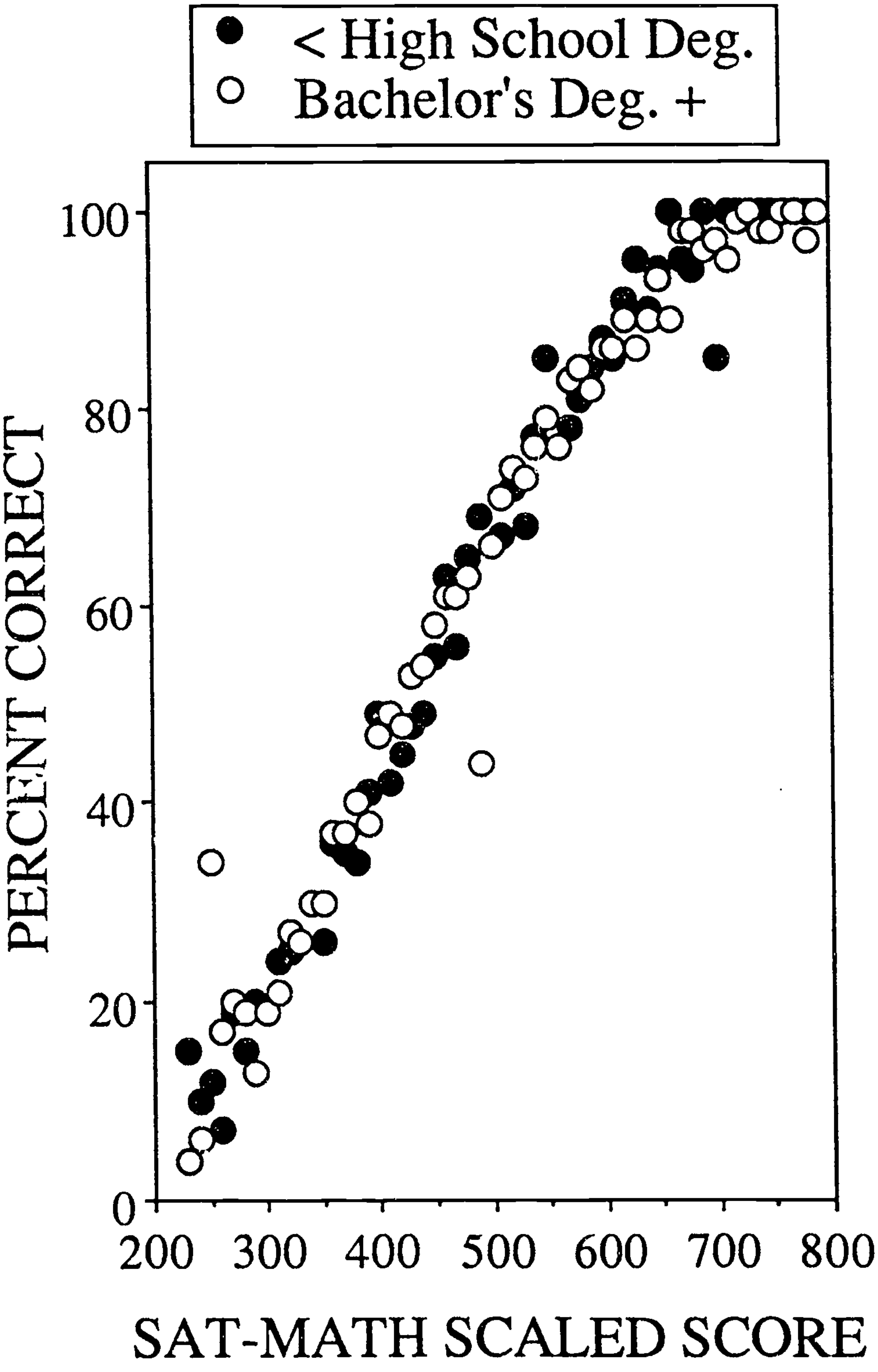
● < High School Deg.
○ Bachelor's Deg. +
100
80
60
40
20
0
PERCENT CORRECT
200
300
400
500
600
700
800
SAT-MATH SCALED SCORE

Figure 3

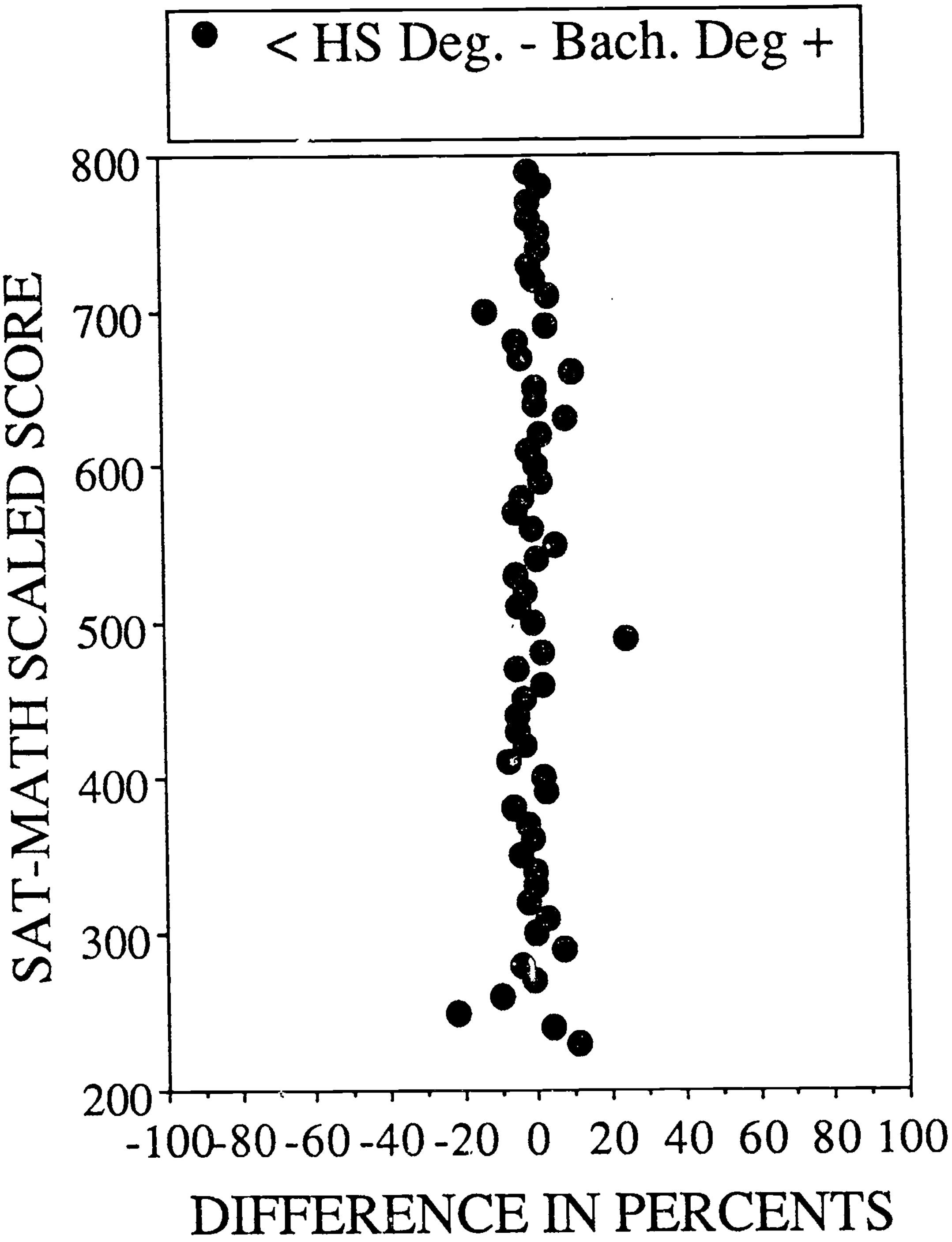
● < HS Deg. - Bach. Deg +
800
700
600
500
400
300
200
SAT-MATH SCALED SCORE
-100-80-60-40-20 0 20 40 60 80 100
DIFFERENCE IN PERCENTS

Figure 4

In contrast, the plots in Figures 3 and 4 depict an item that exhibits negligible DIF. This item came from the second study that used standardization on the SAT (Kulick & Dorans, 1983), in which father's level of education was used to compare groups from different socioeducational groups. Examinees whose fathers had not completed high school (the focal group, N = 7, 053) performed on this SAT-Mathematics item in much the same way as students whose fathers has attained at least a bachelor's degree (the reference group, N = 24,910). Whereas, the decoy : duck item had atypical DIF for its test edition, the DIF for this SAT-Mathematics item was more typical of items on that March 1980 form of the SAT.

3.1 Standardization's Item Discrepancy Indices

The sheer volume of the SAT item pool precludes sole reliance on item-test regression plots and difference plots for DIF assessment. There is a clear need for a numerical index that targets items like that depicted in Figures 1 and 2 for closer scrutiny, while allowing items such as that depicted in Figures 3 and 4 to pass swiftly through the screening process. Standardization has two such flags : the standardized p-difference (STD P-DIF) and the root-mean-weighted squared difference (RMWSD). Both indices use a weighting function supplied by the standardization group to average differences (or squared differences) across levels of the matching variable. The function of the standardization group, which may be a real group or a hypothetical group, is to supply a set of weights, one for each score level, for use in weighting each individual D_m (or D_m^2) before accumulating these weighted differences (or squared differences) across score levels to arrive at a summary item-discrepancy index.

3.1.1 *The standardized p-difference.* The standardized p-difference is defined as:

$$\text{STD P-DIF} = \Sigma_m w_m(E_{fm}-E_{rm})/\Sigma_m w_m = \Sigma_m w_m D_m/\Sigma_m w_m ,$$

where $(w_m/\Sigma w_m)$ is the weighting factor at score level m supplied by the standardization group to weight differences in item performance between the focal group (E_{fm}) and the reference group (E_{rm}). The standardized p-difference is so-named because the original applications of the standardization methodology defined expected item score in terms of proportion correct at each score level,

$$\text{STD P-DIF} = \Sigma_m w_m(P_{fm}-P_{rm})/\Sigma_m w_m = \Sigma_m w_m D_m/\Sigma_m w_m ,$$

where P_{fm} and P_{rm} are the proportions correct, number of examinees who answer correctly over total number of examinees, in the focal and reference groups at score level *m*,

$$P_{fm} = R_{fm}/N_{fm}\,; \qquad P_{rm} = R_{rm}/N_{rm}\,.$$

In contrast to impact, in which each group has its relative frequency serve as a weight at each score level,

$$\text{IMPACT} = P_f - P_r$$

$$= \Sigma_m N_{fm} P_{fm}/\Sigma_m N_{fm} - \Sigma_m N_{rm} P_{rm}/\Sigma_m N_{rm}\,,$$

STD P-DIF uses a standard or common weight on both P_{fm} and P_{rm}, namely, $(w_m/\Sigma w_m)$. The use of the same weight on both P_{fm} and P_{rm}, or more generally E_{fm} and E_{rm}, is the essence of the standardization approach. In the equation above P_r is proportion correct observed in the reference group, while P_f is the proportion correct observed in the focal group.

The particular set of weights employed for standardization depends upon the purposes of the investigation. Some plausible options are the following:

- $w_m = N_{tm}$, the number of examinees at m in the total group;
- $w_m = N_{rm}$, the number of examinees at m in the reference group;
- $w_m = N_{fm}$, the number of examinees at m in the focal group;

or - w_m = the relative frequency at m in some reference group.

In practice, $w_m = N_{fm}$ has been used because it gives the greatest weight to differences in P_{fm} and P_{rm} at those score levels most frequently attained by the focal group under study. Use of N_{fm} means that STD P-DIF equals the difference between the observed performance of the focal group on the item and the predicted performance of selected reference group members who are matched in ability to the focal group members. This can be derived very simply,

$$\text{STD P-DIF} = \Sigma_m N_{fm}(P_{fm}-P_{rm})/\Sigma_m N_{fm}$$

$$= \Sigma_m N_{fm} P_{fm}/\Sigma_m N_{fm} - \Sigma_m N_{fm} P_{rm})/\Sigma_m N_{fm},$$

$$\text{STD P-DIF} = P_f - P_f^*,$$

group predicted from the reference-group item-test regression curve, P_{rm}, or as suggested above, the *predicted* performance of selected *reference* group members who are matched in ability to the focal group.

STD P-DIF is an index that can range from -1 to +1 (or -100% to 100%). Positive values of **STD P-DIF** indicate that the item favors the focal group, while negative **STD P-DIF** values indicate that the item disadvantages the focal group. **STD P-DIF** values between -.05 and +.05 are considered negligible. **STD P-DIF** values between -.10 and -.05 and between .05 and .10 are inspected to insure that no possible effect is overlooked. Items with **STD P-DIF** values outside the {-.10, +.10} range are more unusual and should be examined very carefully.

A delta metric version of the **STD P-DIF** index is:

$$\mathrm{STD\ D\text{-}DIF} = -2.35\ln\{[P_f^*/(1 - P_f^*)]/[P_f/(1 - P_f)]\}\ .$$

STD D-DIF tends to have a smaller variance than **MH D-DIF** across items, and correlates higher with **MH D-DIF** than does **STD P-DIF** across items.

3.2 Standard Errors for Standardization's DIF Indices

The standard errors for the standardization method DIF indices were also developed by Holland. The standard error for the focal group weighting version of **STD P-DIF** is

$$\mathrm{SE(STD\ P\text{-}DIF)} = \{P_f(1\text{-}P_f)/N_f + \mathrm{VAR}(P_f^*)\}^{.5}\ ,$$

where,

$$\mathrm{VAR}(P_f^*) = [\Sigma_m N_{fm}{}^2 P_{rm}(1\text{-}P_{rm})/(N_{rm}N_f{}^2)]\ .$$

The standard error for the focal group weighting version of **STD D-DIF** is

$$\mathrm{SE(STD\ D\text{-}DIF)} = (2.35)\{[(P_f(1\text{-}P_f)N_r)^{-1} + \mathrm{VAR}(P_f^*)/P_f^*(1 - P_f^*)\}^{.5}\ ,$$

where N_r is the number of examinees in the reference group.

3.3 Differential Distractor Functioning, Speededness and Omission

DIF assessment does not stop with the flagging of an item for statistical DIF. In fact, the flagging step can be viewed as just the beginning. The next step is to try to understand the reason or reasons for the DIF. Green, Crone, and Folk (1989) have developed a log-linear approach for assessing what they call *differential distractor functioning* (DDF). The standardization approach to distractor analysis can also be quite helpful. Some of the items identified by Green, Crone and Folk will be analyzed from the standardization framework below; some of these items are also analyzed in Thissen, Steinberg and Wainer (in press) for *differential alternative functioning (daf).*

3.3.1 *Differential Distractor Functioning.* The generalization of the standardization methodology to all response options including omission and not reached is straightforward and is known as standardized distractor analysis (Dorans, Schmitt, & Bleistein, 1988, 1989). It is as simple as replacing the keyed response with the option of interest in all calculations. For example, a standardized response rate analysis on option **A** would entail computing the proportions choosing **A** (as opposed to the proportions correct) in both the focal and reference groups,

$$P_{fm}(A) = A_{fm}/N_{fm}; \quad P_{rm}(A) = A_{rm}/N_{rm} ,$$

where A_{fm} and A_{rm} are the number of people in the focal and reference groups, respectively, at score level m who choose option **A**. The next step is to compute differences between these proportions,

$$D_m(A) = P_{fm}(A) - P_{rm}(A).$$

Then these individual score level differences are summarized across score levels by applying some standardized weighting function to these differences to obtain STD P-DIFF(A),

$$\text{STD P-DIF}(A) = \Sigma_m w_m D_m(A)/\Sigma_m w_m ,$$

the standardized difference in response rates to option **A**. In a similar fashion one can compute standardized differences in response rates for options **B**, **C**, **D**, and **E**, and for non-responses as well.

The plots produced by the standardized distractor analyses can be quite helpful in trying to interpret DIF data. As an example, consider the plots in Figure 5. Portrayed are selected empirical option response curves for an SAT antonym item from a disclosed 1984 test form for which the key, distractors and DIF information are provided below:

STD P-DIF (Option)

MA	PR	BLK	PRACTICAL:	
4	9	**12**	(A)	difficult to learn
0	0	0	(B)	inferior in quality
1	1	1	(C)	providing great support
-5	-11	-16	(D)	having little usefulness
0	0	0	(E)	feeling great regret

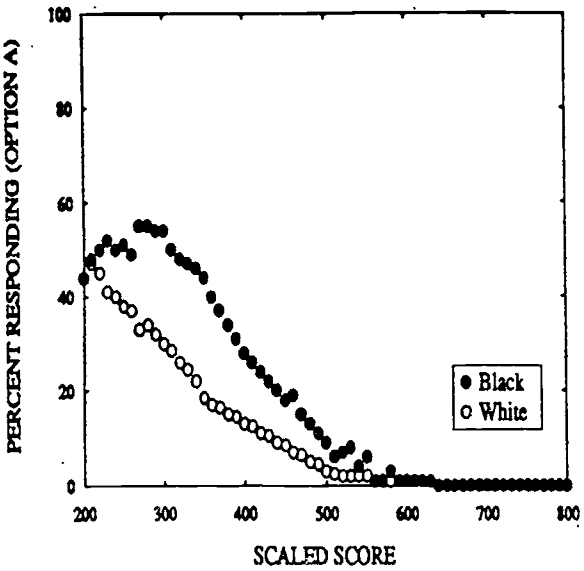

Figure 5a

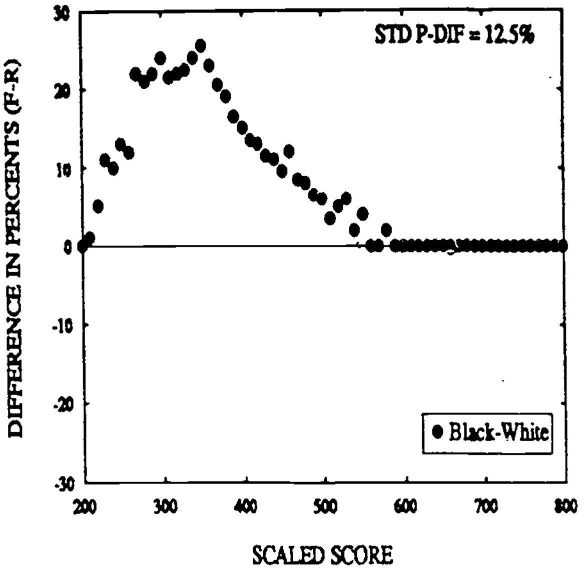

Figure 5b

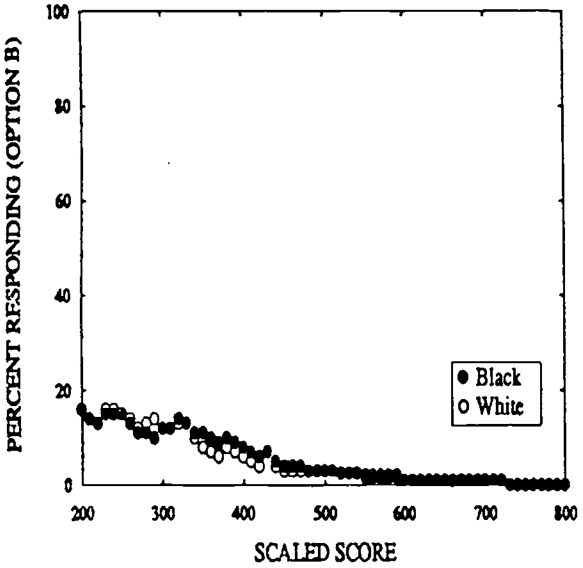

Figure 5c

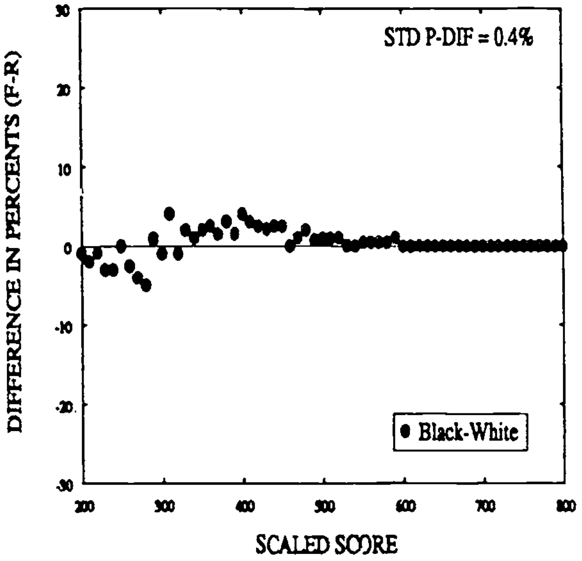

Figure 5d

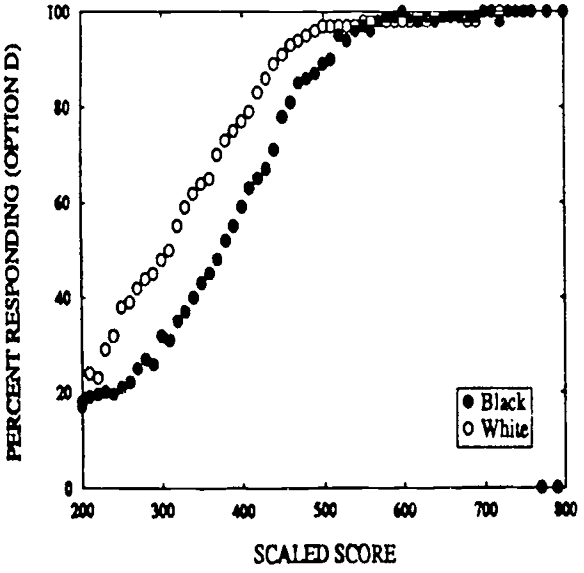

Figure 5e

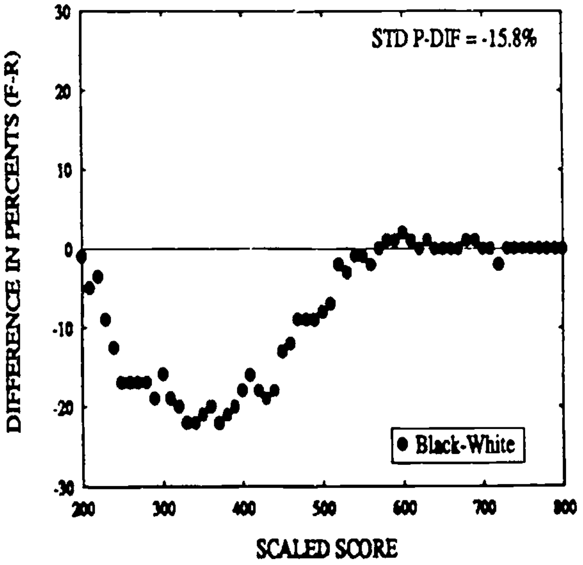

Figure 5f

As can be seen in the table, standardization identifies DIF on the key, the opposite of *practical* is *(D) having little usefulness*, for Blacks (BLK **STD P-DIF** = -16%) and Puerto Ricans (PR **STD P-DIF** = -11%), but only marginally for Mexican Americans (MA **STD P-DIF** = -5%). In addition, the **STD P-DIF(option)** values indicate where the "anti-DIF" may lie, and the plots for the Black group corroborate these indications. Clearly, the Black and Puerto Rican focal groups are drawn towards *(A) difficult to learn*, which suggests that they have confused the word *practical* with the word "practice".

For additional examples, we will use the two SAT items reported by Green, Crone and Folk (1989) to exhibit relatively small differential distractor functioning, and substantial differential distractor functioning. The standardized distractor information for the item with relatively small differential distractor functioning is shown below:

STD P-DIF (Option)				
AA	HISP	BLK	DECADENT:	
-1	0	1	(A)	enormously wealthy
-4	-1	-1	(B)	remarkably charming
-3	-2	-2	(C)	ruthless
-1	0	0	(D)	distinctive
5	5	3	(E)	flourishing
.63	.76	.44	MH D-DIF	

This item exhibits marginal positive DIF for all three focal groups, Asian Americans (AA), Hispanics (HISP), and Blacks (BLK)., Likewise, there is very little differential distractor functioning, as measured by the standardization method.

The data for the second item identified by Green, Crone and Folk (1989) is more interesting:

In some animal species, differences between opposite sexes are so ____ that it is difficult to tell that the male and female are _____.

STD P-DIF (Option)				
AA	HISP	BLK		
2	**10**	3	(A)	measurable . . . . distinct
1	-2	-3	(B)	minute. . . . similar
2	2	4	(C)	obvious . . . .indistinguishable
-8	**-12**	-9	(D)	extreme . . . . related
2	1	3	(E)	trivial . . . . identical
-1.09	-1.40	-1.05	MH D-DIF	

There is a moderate level of DIF for all three focal groups on this item; for Hispanics, the DIF level is particularly noticeable. The standardized distractor information is particularly informative for this focal group, who are drawn towards option *(A) measureable . . . . distinct* in much greater proportions than the matched group of whites, who are drawn to option *(D) extreme . . . . related*. It is not clear to us why Hispanics are drawn towards (A), nor why all three focal groups exhibit negative DIF on this item. While the distractor analysis tells where the "anti-DIF" is, it doesn't tell us why it's there. See Thissen, Steinberg, and Wainer (in press) for a *daf* analysis of this item. See Schmitt, Holland and Dorans (in press) for examples in which the standardized distractor analysis corroborates DIF hypothesis for Hispanics.

3.3.2 *Differential Speededness.* Application of the standardization methodology to counts of examinees at each level of the matching variable who did not reach the item results in a standardized not-reached difference,

$$\text{STD P-DIF(NR)} = \Sigma_m w_m (P_{fm}(NR) - P_{rm}(NR)) / \Sigma_m w_m \,.$$

For items at the end of a separately-timed section of a test, these standardized differences provide measurement of the differential speededness of a test. Differential speededness refers to the existence of differential response rates between focal group members and matched reference group members to items appearing at the end of a section. Schmitt and Bleistein (1987) found evidence of this phenomenon for Blacks, as compared to a matched group of Whites, on analogy items. Schmitt and Dorans (1990) reported that this effect was also found for Hispanics. In Dorans, Schmitt and Bleistein (1988), differential speededness results for Black, Hispanic and Asian-American focal groups, compared to a White reference group, are presented and their implications are discussed. In Dorans, Schmitt and Curley (1988), the effects of item position on differential speededness and on DIF assessment were investigated. This study, which is described in more detail in Schmitt, Holland and Dorans (in press), found that excluding examinees who do not reach an item from the calculation of the DIF statistic for that item partially compensates for the effects of item location on the DIF estimate.

One implication that the existence of differential speededness has for analyzing DIF or DDF is that the matching variable, total score, may be contaminated due to differential speededness. Research presently being conducted by A. Schmitt and her colleagues may shed light on the seriousness of this potential contamination and the efficacy of potential solutions to the problem, such as matching on a shortened unspeeded portion of the total test. Simulation studies should prove useful here.

3.3.3 *Differential Omission.* It should be obvious that standardization can also be applied to the study of differential omission. In fact, Schmitt and Dorans (1990) report on some of these studies including one by Rivera and

Schmitt (1988) who found that while Hispanics as a group omit more than Whites on the SAT, Hispanics tend to omit less than whites of comparable ability. This is a clear example of Simpson's paradox in terms of omitting behavior, an example which had immediate implications for the type of advice that was being offered to Hispanic test-takers. On the basis of the marginal distributions, it appeared that Hispanics were omitting less than Whites. After conditioning on total test score, it became clear that the opposite was true. So we close our discussion of the Mantel Haenszel and standardization methods with another illustration of the need to compare the comparable.

4. Mantel-Haenszel and Standardization From a Common Framework

Up to now, the Mantel-Haenszel method and the standardization method have been described from the the frameworks from which they evolved: Mantel-Haenszel as a powerful statistical test of the constant odds ratio model, and standardization as a non-parametric alternative to item response theory for describing item-ability regressions. The two procedures, however, share a common framework spelled out in Dorans (1989).

For rights-scored tests, the standardization definition of null DIF is in terms of zero p-differences at all levels of the matching variable,

$$R_{fm}/N_{fm} - R_{rm}/N_{rm} = 0 \qquad m = 1, ..., M .$$

The definition of null DIF for Mantel-Haenszel is

$$[R_{rm}/W_{rm}] / [R_{fm}/W_{fm}] = 1 \qquad m = 1, ... , M .$$

When null DIF holds, the standardization definition can be rearranged as:

$$R_{fm}/N_{fm} = R_{rm}/N_{rm} ,$$

$$R_{fm}N_{rm} = R_{rm}N_{fm} ,$$

$$R_{fm}(W_{rm} + R_{rm}) = R_{rm}(W_{fm} + R_{fm}) ,$$

$$R_{fm}W_{rm} + R_{fm}R_{rm} = R_{rm}W_{fm} + R_{fm}R_{rm} ,$$

$$R_{fm}W_{rm} = R_{rm}W_{fm} ,$$

$$R_{rm}/W_{rm} = R_{fm}/W_{fm} ,$$

which becomes the Mantel-Haenszel definition of null DIF,

$$[R_{rm}/W_{rm}] / [R_{fm}/W_{fm}] = 1 \qquad m = 1,\ldots, M .$$

Mantel-Haenszel and standardization share a common definition of null DIF that is stated in different metrics. The two procedures differ with respect to how they measure departures from null DIF.

Under rights scoring for the items in which responses are coded as either correct or incorrect (including omissions), both the standardization procedure and the Mantel-Haenszel procedure use the same basic data to focus on differences in conditional item performance, which can be operationalized as differences in non-parametric item test regressions (standardization) or in terms of a constant odds ratio model (Mantel-Haenszel). As we have seen earlier, counts of rights and wrongs on each item can be arranged into a **2(groups)-by-2(item scores)-by-M(score levels)** contingency table for each item being studied.

The Mantel-Haenszel and standardization procedures operate on the basic data of the **2(groups)-by-2(item scores)-by-M(score levels)** contingency table in different ways. As a consequence, they measure departures from the null DIF condition in slightly different ways.

The first difference in how the two procedures measure departures from null DIF is in the metric for defining DIF. Standardization uses differences in conditional proportions correct,

$$D_m = P_{fm} - P_{rm},$$

while Mantel-Haenszel uses conditional odds ratios,

$$\alpha_m = [R_{rm}/W_{rm}]/[R_{fm}/W_{fm}] = [R_{rm}W_{fm}]/[R_{fm}W_{rm}] .$$

The second difference in DIF measurement is in the choice of weights used to average the D_m or the α_m across levels of the matching variable. The Mantel-Haenszel approach uses weights that are nearly optimal statistically for testing a constant odds-ratio model. These weights are:

$$MH_m = W_{rm}R_{fm}/N_{tm} ,$$

such that

$$\alpha_{MH} = \Sigma_m MH_m \; \alpha_m / \Sigma_m MH_m .$$

In contrast, the weights employed in the standardization approach are not defined statistically. Instead they may be chosen to suit the needs of a particular investigator. This flexibility has not be utilized often. Instead the intuitively appealing focal group frequency distribution, which was employed by Dorans and Kulick (1983) in their original work on the SAT, is typically used to describe departures from null DIF,

$$\text{STD P-DIF} = \Sigma_m N_{fm}(P_{fm}-P_{rm})/\Sigma_m N_{fm}\ .$$

Holland and Thayer (1988) pointed out that Cochran (1954) developed a set of weights for the p-difference metric that are statistically motivated, that is they are appropriate for testing a constant difference model across score levels., These weights are:

$$C_m = N_{rm}N_{fm}/N_{tm}\ .$$

The third difference in DIF measurement between the two methods is the metric in which the final statistic is portrayed. Although, a delta metric version of the standardization DIF statistic has been developed, the primary, almost exclusive, metric used by standardization has been the p-metric, even for formula-scored tests where an item formula-scored metric would seem superior on logical grounds. In contrast, the delta metric has been the metric of choice for the Mantel-Haenszel method. One consequence of this difference in choice of metrics is that standardization tends to down play DIF in easy and hard items because the p-metric is bounded at both the top and bottom. In contrast, the delta metric is unbounded at the extremes, and consequently differences for easy and hard items are played up

Despite these differences in choice of metric and weighting, standardization and Mantel-Haenszel agree very closely with respect to measurement of departures from null DIF for the vast majority of items. In fact, correlations across items between the two DIF methods in the same metric, e.g. delta, are typically close to unity, and slightly higher than within-method correlations between metrics, which are in the high nineties. Cross-metric cross-method correlations across items are usually in the mid-nineties. These correlations indicate that the two methods are measuring essentially the same thing, DIF, in slightly different ways; intuitively appealing weighting of conditional differences in proportions correct vs. statistically-driven weighting of conditional odds ratios. The correlations also indicate that the choice of metric for describing the DIF effect may be more critical from a practical point of view than the choice of method.

5. Implementation Issues

DIF implementation at ETS occurred quickly once Mantel-Haenszel was selected as the method for DIF detection and standardization was selected as the method for DIF description. With implementation came an assortment of issues that required temporary if not permanent resolution. In this section of the report, several of these issues are discussed. In the next section, future research associated with issues that remain either unsolved or only partially solved are discussed.

5.1 Inclusion of Studied Item

Holland and Thayer's (1988) analysis of the interrelationship between Mantel-Haenszel and the Rasch model led to some counterintuitive conclusions about whether an item should be included as part of the criterion when DIF analysis is performed on the item. Holland and Thayer (1988) concluded on theoretical grounds that an item should be included as part of the matching variable:

> *If it is not included, then the MH procedure will not behave correctly when there is no dif according to an IRT model. However the Rasch analysis suggests that the inclusion of the studied item in the matching criterion does not mask the existence of dif, rather it is the inclusion of other items exhibiting dif in the criterion that could lead to the finding that no dif exists for the studied item when in fact it does.* (p. 141)

The need to ensure that other items in the matching criterion are free of DIF is one argument for criterion refinement, a procedure described below.

The mathematical argument for inclusion of the item in the matching variable is presented in Holland and Thayer (1988), who also show how trivial it is to correct the M 2-by-2 tables for rights-scored tests in which number right score is the matching criterion. The correction for formula-scored tests, however, is not so trivial.

5.2 Criterion Refinement or Purification

An argument that is often voiced against the Mantel-Haenszel procedure, the standardization procedure, and other DIF assessment techniques that use an internal criterion is the circularity involved in using total test score as a criterion for matching. Although not a perfect matching criterion because all tests contain a certain amount of statistical noise, scores on a test are often the best available matching criterion for several reasons. First, the total test score is often a much more reliable measure of what any individual item purports to measure. Second, many test scores have demonstrated validity for their intended purposes. Third, test scores are typically obtained under the same conditions for all examinees.

Despite these advantages of reliability, validity, and standardized administration, tests scores are criticized because items are part of the score, and there is a concern about the circularity of using potentially biased test scores as a criterion for DIF analyses. The most direct way of demonstrating that the total test score is acceptable as a matching variable is to demonstrate that it is valid for its intended purposes, and that it is equally valid for all focal and reference groups. DIF analysis is not a substitute for validity studies. In fact, the DIF analysis assumes that the criterion is valid and fair.

Since all tests are imperfect, they may in fact contain some items which do have DIF. Otherwise, the DIF analysis would be a meaningless exercise. In an attempt to ensure that the matching criterion is in fact DIF-free, DIF analyses at ETS occur in two steps. The first step is called the criterion refinement or purification step. Here, items on the matching variable are analyzed for DIF, and any items that exhibit sizeable DIF are removed regardless of the sign of the DIF. Then this refined criterion is used for another DIF analysis of the same items and any other items that were not included in the criterion refinement step.

6. Future Directions

DIF implementation is in a nascent stage. Much basic research has been done, but much more needs to be done. Our methodologies for DIF assessment are good, but could be better. In this section, areas for further methodological research are identified. These areas fall into three major classes: the matching variable; the studied variable; and the group variable.

6.1 The Matching Variable

6.1.1 *Dimensionality and DIF: The need for multivariate matching.* Items with sizable DIF are items that behave differently for one group. This difference indicates that the identified item does not appear to measure the same construct as the total test. Thus DIF measures violations from unidimensionality. The unidimensionality of the matching variable is central to the DIF assessment process. Shepard (1982) stresses this by saying:

> *...it should be clear that the assumption of unidimensionality underlying all of the the (DIF) methods is not merely a statistical prerequisite but it is central to the way in which item bias is defined.* (p. 25)

Later, Shepard (1987) discusses how multidimensionality and DIF interact:

> *It is also generally understood that the various (DIF) procedures function by signalling multidimensionality. Therefore, the statistical indices can detect when subparts of the test are measuring differently for different groups, but are not automatically evidence of bias. To address the issue of bias requires re-examination of the original construct; is the source of multidimensionality some irrelevant difficulty (hence bias) or a valid subdimension of the intended construct.* (p. 1)

From a factor analytic point of view, multidimensionality abounds in item data. Each item is a measure of what the total test measures, i.e. what it has in common with other items, and what it alone measures, its unique item factor. When a test is composed of unidimensional items, as is the case for the mathematical portion of the Scholastic Aptitude Test, DIF occurs when subgroup differences along the unique item dimension do not reflect subgroup differences in developed mathematical ability. When a test is measuring multiple dimensions, as is likely to be the case with a science achievement test, DIF may reflect unique item factor differences between subgroups or the fact that subgroups vary in different ways on the different dimensions measured by the test. DIF is a violation of unidimensionality, but simple interpretation of DIF requires a unidimensional matching variable. See Bleistein and Schmitt (1989), Dorans and Schmitt (1989), Hu and Schmitt (1989), Mazzeo (1989), and Morgan (1989) for a series of papers on the interplay between DIF assessment and dimensionality.

A multidimensional matching variable complicates DIF assessment. Multivariate matching, however, may provide a solution to the problem of multidimensionality. In multivariate matching, examinees are matched on more than one variable. For example, a general developed ability test might be composed of verbal reasoning and mathematics items. Matching on a total score might reveal that the verbal items exhibit positive DIF for females, while the mathematics items exhibit negative DIF. One option is to perform separate DIF analyses for the verbal items and for the mathematics items, as is now done with the SAT. Another option is to match on both the verbal score and the mathematics score prior to comparing how the items function in both groups.

Multivariate matching can have heavy data requirements because of need to cross the levels of all the variables that go into the match. In addition, data may be sparse for many combinations of the two or more variables especially if they are highly correlated. Where data are sparse, separate analyses against more unidimensional criteria, e.g. math items against a math score, and verbal items against a verbal score, may be the only practical option. Methods such as propensity score matching (Rosenbaum & Rubin, 1985) may be a useful solution when data are sparse.

6.1.2 *Inclusion of studied item for formula-scored tests.* As mentioned earlier, Holland and Thayer (1988) demonstrated how easy it is to adjust the MH calculations for inclusion of the studied item in the matching criterion when the matching criterion is a number-right score. Inclusion of the studied item with a formula scored criterion is not at all straightforward because it is not a simple matter to adjust the matching variable after the formula score has been rounded to integer format. As a consequence, some peculiar practices have evolved with DIF analyses for formula-scored tests. For example, analyses of studied items that are external to the matching variable, e.g. pretest items collected in the non-operational section of the SAT, are done against a rights-scored criterion despite the fact that the test was administered under formula-scored conditions. Under formula scored conditions, omitting an item is different than getting it incorrect. Under rights scored conditions, omitting an item is treated the same as getting the item incorrect. So in order to include the pretest item in the criterion, the matching variable is scored in a manner that is inconsistent with test administration conditions.

One potential solution to this problem is to employ multivariate matching on rights, wrongs, omits and not reached (presently examinees who do not reach the item are excluded from the calculation of the DIF statistic). Another option is to use a version of formula scoring in which a correct response is assigned a score equal to the number of response options, an omit or not reached is assigned a one, and a wrong is assigned a zero. Under this type of formula scoring, there are no fractions, and hence no need to round to integer format. Hence the adjustment for inclusion may be as simple as it is for rights scored tests.

6.2 Studied Variable

6.2.1 *Formula score DIF.* It has been a common practice to rights score items for the purpose of item analysis regardless of the conditions under which the item was administered. For rights scored tests, this is a perfectly reasonable practice. For formula-scored tests, however rights scoring of the item is not consistent with the conditions under which the item was administered. Had the examinee known an item was to be rights scored, it is unlikely he would have omitted that item since omitting is tantamount to getting the item wrong on a rights scored test.

The DIF computer programs used at ETS employ rights scoring of items for both Mantel-Haenszel and standardization to obtain MH D-DIF, MH P-DIF, STD P-DIF, and STD D-DIF. In addition, the program can be asked to compute a little-used standardization statistic, which may in fact be the best standardization statistic to use for formula-scored tests, such as the SAT,

$$\text{STD FS-DIF} = \Sigma_m w_m(E_{fm}-E_{rm})/\Sigma_m w_m = \Sigma_m w_m D_m/\Sigma_m w_m \,,$$

where instead of scoring the item 1 if correct and 0 if incorrect or omit, which yields **STD P-DIF**, the item is scored 1 if correct, 0 if omit, and -1/(k-1) if incorrect, where k is the number of response options. Under this type of scoring, the expected item performance in the focal group at score level **m** is

$$E_{fm} = \{R_{fm}*(1) + O_{fm}*(0) + W_{fm}*(-1/(k-1))\}/N_{fm},$$

where R_{fm}, W_{fm}, and O_{fm} are counts of the number right, the number wrong, and the number of omits, respectively at score level **m** in the focal group and N_{fm} is the sum of R_{fm}, W_{fm}, and O_{fm}. Likewise, for the reference group, we have

$$E_{rm} = \{R_{rm}*(1) + O_{rm}*(0) + W_{rm}*(-1/(k-1))\}/N_{rm},$$

Unlike **STD P-DIF, STD FS-DIF** does not range from -1 to +1. Instead, its theoretical range is -k/(k-1) to +k/(k-1). Under no omitting, which is likely for easy items, STD FS-DIF = k/(k-1) STD P-DIF. These two standardization indices are more likely to diverge when items are difficult and omitting becomes a dominant behavior.

6.2.2 *Testlet DIF*. Most DIF assessment procedures are just that differential item functioning procedures; the item is the unit of analyses. Some differential functioning issues are better answered at a larger level of analysis, such as performance on a set of reading passage items, or performance on a set of items of comparable content. Here, the unit of analyses shifts to the testlet (Wainer & Kiely, 1987). Special types of testlets called item parcels have been useful in dimensionality assessment (Dorans & Lawrence, 1987). Wainer and Lewis (1990) have shown other areas where testlet-level analysis has also proved superior to item-level analysis. Dorans and Lawrence (1987) argue that parcel (testlet) analysis may be preferable to item analysis because testlets are more reliable indicators than single items. There exists a need to develop and try out procedures for Testlet DIF, or DTF to be exact. Some promising possibilities are the flexible standardization method (Dorans & Kulick, 1986), IRT-based models developed by Thissen and his colleagues and linear regression procedures.

The standardization approach could be readily adapted to testlet DIF by replacing expected item performance with expected testlet performance in the basic standardization equation. This would result in a comparison of empirical testlet-test regressions, using a standard weighting function to produce numerical indices that describe how far apart these regressions are for some standardization population.

As the number of items in a testlet increases, the more likely it is that the testlet-test regression will be linear provided item difficulties are somewhat spread out among items defining the testlet. In that case, a comparison of linear regressions would be possible.

6.3 Group Variable

6.3.1 *Melting-pot DIF*. Hu and Dorans (1989) recently found that removal of an item that was flagged for positive gender DIF lowered females scores slightly and raised males scores slightly, as expected. It also had some unintended consequences. It raised the scores of Hispanics and Asian-Americans more than it raised male scores. In addition, it raised the scores of Hispanic and Asian-American females despite the fact that deletion of this item with positive gender DIF reduced the overall female mean score. In addition to pointing out that deleting items for DIF can have unintended consequences for the groups that were not the focus of analysis, this finding demonstrates a flaw with the "marginal DIF analysis" that we do now. Instead of crossing gender with ethnicity/race to study DIF, we look at the margins, i.e., we do DIF analyses on gender and we do DIF analyses on ethnicity/race. This "marginal DIF analysis" ignores potential interactions between gender and ethnicity/race, interactions that may be important. One possible solution to this problem is to do Melting-Pot DIF analyses in which the reference group is the population of all test-takers who meet the appropriate grade level and language proficiency criteria, the melting pot group. Each gender/ethnic group is a focal group. Melting-Pot DIF would permit one to do gender comparisons within ethnic group, as well as ethnic group comparisons with gender group. Marginal DIF analyses could be obtained, of course, by collapsing across the other margin. One advantage of Melting-Pot analysis is that everybody is a focal group member once and a reference group member once. Ahother advantage is that more DIF is more likely to be found in the smaller subpopulations because they are a smaller part of the melting pot. On the negative side, DIF will be harder to find in the larger groups, such as White females and White males. One possible solution to the problem would be to borrow Wainer's (1989) notions of *standardized impact*, similar to the standardization index, and *total standardized impact*, which can be obtained by weighting the *standardized impact* by the number of individuals in the focal group. This practice, however, might introduce the opposite problem: small groups would be ignored.

6.3.2 *Educational Advantage Construct*. As DIF implementation moves swiftly along at ETS and elsewhere, it is clear that several fundamental issues require more attention. Several of these issues have been discussed in this section of the report. One very important issue that remains to be discussed is that of focal group definition. To date, focal groups have been intact easily-defined groups such as Asians, Blacks, Hispanics and females. References groups have been Whites or males. It could be argued, however, that these intact ethnic groups are merely surrogates for an educational disadvantage attribute that should be used in focal group definition. In fact, within any of these groups, there is probably a considerable degree of variability with respect to educational advantage or disadvantage. Perhaps, we should be focusing our group definition efforts towards defining and measuring educational advantage or disadvantage directly. This argument echoes that made more than a decade ago in the American Psychologist by Novick and Ellis (1977), where a strong case was made for "the explicit identification of those attributes that constitute disadvantage, rather than accepting group membership as a surrogate for disadvantage" (p. 318), and more recently by Schmitt and Dorans (1990). Novick and Ellis acknowledged that the problems of understanding what constitutes disadvantage and being able to measure it adequately were formidable. They still are. Significant advances in DIF implementation, however, may depend on serious efforts that address this issue.

7. Closing Comments

The major purpose of this report was to present the Mantel-Haenszel technique for DIF detection and the standardization technique for DIF description. We began by making the important distinction between DIF and Impact, pointing the need to compare the comparable. Then the Mantel-Haenszel procedure and the standardization procedure were described in some detail in that order. A common framework was used to present similarities and dissimilarities between the two methods. Then we discussed relationship of the MH procedure to IRT methods for DIF detection in general, and the Rasch model, in particular. Then the use of standardization for assessing differential distractor functioning, differential speededness and differential omission was presented.

Several issues in applied DIF analyses were discussed including, inclusion of the studied item in the matching variable, and the refinement of the matching variable. Future research topics dealing with the matching variable, the studied variable and the group variable were discussed.

Large scale DIF implementation is a relatively new phenomenon in the field of measurement. Low-cost, practical, statistically-sound techniques, like the Mantel-Haenszel and standardization approaches, have made large scale implementation a reality. These are powerful techniques for DIF detection and description. As the implementation issues and future direction sections of this report indicate, these procedures could be improved, made more applicable to the actual testing situation. Although they are sound methods for DIF assessment, enhancements can and should be made. The major focus of future DIF research efforts, however, should not be on methodological enhancements. Although it could be improved, the methodology is quite sound. Future research should focus on trying to uncover testable, verifiable, robust explanations for why DIF occurs when it does. As Schmitt, Holland and Dorans (in press) reveal, this will not be an easy task, partly because DIF is usually small relative to other item properties such as difficulty, partly because DIF research is constrained by many practical and ethical constraints, and partly because DIF, like bias, is a political issue, as well as an issue that is laden with emotional overtones. The major challenge facing the DIF field is to take the methods described in this report or the methods described in Thissen, Steinberg and Wainer (in press) and use them to identify replicable DIF, generate sound hypotheses about this replicable DIF, test these hypotheses under controlled conditions, and develop guidelines for producing future tests that are free from these irrelevant sources of group differences.

References

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (RR-81-16). Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Ford, S. F. (1973). Item-race interactions on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95-105.

Bleistein, C. A., & Schmitt, A. P. (1989, March). *Criterion selection for evaluation of DIF for Chemistry Achievement tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67,* 364-365.

Cochran,W. G. (1954). Some methods for strengthening the common χ^2 test. *Biometrics, 10,* 417-451.

Dorans, N. J. (1982). *Technical review of item fairness studies: 1975-1979* (SR-82-90). Princeton, NJ: Educational Testing Service.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217-233.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (RR-83-9). Princeton, NJ: Educational Testing Service.

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

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (RR-87-35). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Schmitt, A. P. (1989, March). *The methods for dimensionality assessment and DIF detection.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speededness* (RR-88-31). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1989). *The standardization approach to differential distractor functioning: Assessing differential speededness*. Manuscript under review.

Dorans, N. J., Schmitt, A. P., & Curley, W. E. (1988, April). *Differential speededness: Some items have DIF because of where they are, not what they are*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26,* 147-160.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th Annual Conference of the Military Testing Association*, San Diego, CA, *Vol. 1, pp. 282-287.*

Holland, P. W. (1989). A note on the covariance of the Mantel-Haenszel log-odds estimator and the sample marginal rates. *Biometrics, 45,* 1009-1015.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (RR-85-43). Princeton, NJ: Educational Testing Service.

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

Hu, P. G., & Γorans, N. J. (1989, March). *The effects of deleting items with extreme differential item functioning on equating functions and reported score distributions*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hu, P. G., & Schmitt, A. P. (1989, March). *Evaluation of matching criteria for differential item functioning analyses of Biology Achievement Tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Kulick, E., & Dorans, N. J. (1983). *Assessing unexpected differential item performance of candidates reporting different levels of father's education on SAT Form CSA2 and TSWE Form E29* (SR-83-27), Princeton, NJ: Educational Testing Service.

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

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mazzeo, J. (1989, March). *A study of the dimensionality of the ATP Achievement test in Chemistry.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Morgan, R. (1989, March). *An examination of the dimensional structure of the ATP Biology Achievement test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Novick, M. R., & Ellis, D. D. (1977). Equal opportunity in educational and employment selection. *American Psychologist, 22,* 306-320.

Phillips, A. & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43,* 425-431.

Ramsey, P. (in press). Sensitivity review process. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rivera, C., & Schmitt, A. P. (1988). *A comparison of Hispanic and White students' omit patterns on the Scholastic Aptitude Test* (RR-88-44). Princeton, NJ: Educational Testing Service.

Robins, J., Breslow, N. E., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42,* 311-324.

Rosenbaum, P. R., & Rubin, D. R. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39,* 33-38.

Scheuneman, J. D. (1975, April). *A new method of assessing bias in test items.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 106 359).

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items* (RR-87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., & Dorans, N. J. (Eds.), (1987). *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27,* 67-81

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (in press). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias.* (pp. 9-30). Baltimore: Johns Hopkins University Press.

Shepard, L. A. (1987). Discussant comments on the NCME symposium, Unexpected Differential Item Performance and its Assessment Among Black, Asian-American, and Hispanic Students. In A. P. Schmitt & N. J. Dorans (Eds.), (1987). *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton, NJ: Educational Testing Service.

Simpson, E. H. (1951). The interpretation of interaction contingency tables. *Journal of Royal Statistical Society* (Series B), *13,* 238-241.

Thissen, D. (1987). Discussant comments on the NCME symposium, Unexpected Differential Item Performance and its Assessment Among Black, Asian-American, and Hispanic Students. In A. P. Schmitt & N. J. Dorans (Eds.), (1987). *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton, NJ: Educational Testing Service.

Thissen,D., Steinberg, L., & Wainer, H. (in press). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician, 36,* 46-48.

Wainer, H. (1986). Minority contributions to the SAT Score turnaround: An example of Simpson's paradox. *Journal of Educational Statistics. 11,* 239-244.

Wainer, H. (in press,). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

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

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrics,* 2, 121-134. (Reprinted in *Statistical Papers of George Udny Yule,* selected by A. Stuart and M. G. Kendall, New York, Hafner, 1971.)