DOCUMENT RESUME

ED 384 657 TM 023 950

AUTHOR Dorans, Neil J.; Schmitt, Alicia P.

TITLE Constructed Response and Differential Item Functioning: A Pragmatic Approach.

INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-91-47
PUB DATE Aug 91
NOTE 52p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Computer Assisted Testing; *Constructed Response; Educational Assessment; *Item Bias; Performance; Racial Differences; Scoring; Test Construction; *Test Items
IDENTIFIERS Paper and Pencil Tests

ABSTRACT

Differential item functioning (DIF) assessment attempts to identify items or item types for which subpopulations of examinees exhibit performance differentials that are not consistent with the performance differentials typically seen for those subpopulations on collections of items that purport to measure a common construct. DIF assessment requires a rule for scoring items and a matching variable on which different subpopulations can be viewed as comparable for purposes of assessing their performance on items. Typically, DIF is operationally defined as a difference in item performance between subpopulations, e.g., Blacks and Whites, that exists after members of the different subpopulations have been matched on some total score. Constructed-response items move beyond traditional multiple-choice items, for which DIF methodology is well-defined, towards item types involving selection or identification, reordering or rearrangement, substitution or correction, completion, construction, and performance or presentation. This paper defines DIF, describes two standard procedures for measuring DIF and indicates how DIF might be assessed for certain constructed-response item types. The description of DIF assessment presented in this paper is applicable to computer-delivered constructed-response items as well as paper and pencil delivered items. (Contains 67 references and 5 tables.) (Author)

RESEARCH REPORT

CONSTRUCTED RESPONSE AND DIFFERENTIAL ITEM FUNCTIONING: A PRAGMATIC APPROACH

Neil J. Dorans
Alicia P. Schmitt

Educational Testing Service
Princeton, New Jersey
August 1991

Constructed Response and Differential Item Functioning: A Pragmatic Approach[1]

Neil J. Dorans and Alicia P. Schmitt

Educational Testing Service

[1]To appear in R. E. Bennett, & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement,* published by Lawrence Erlbaum Associates, Hillsdale, NJ. The opinions expressed in this paper are those of one or both of the authors and should not be misconstrued to represent official policy of either the College Board or Educational Testing Service. The authors thank Randy Bennett and William Ward for their reviews and comments on an earlier version of the paper and Elise Sharrett for her assistance in revising the paper.

Abstract

Differential item functioning (DIF) assessment attempts to identify items or item types for which subpopulations of examinees exhibit performance differentials that are not consistent with the performance differentials typically seen for those subpopulations on collections of items that purport to measure a common construct. DIF assessment requires a rule for scoring items and a matching variable on which different subpopulations can be viewed as comparable for purposes of assessing their performance on items. Typically, DIF is operationally defined as a difference in item performance between subpopulations, e.g., Blacks and Whites, that exists after members of the different subpopulations have been matched on some total score. Constructed-response items move beyond traditional multiple-choice items, for which DIF methodology is well-defined, towards item types involving selection or identification, reordering or rearrangement, substitution or correction, completion, construction, and performance or presentation. This paper defines DIF, describes two standard procedures for measuring DIF and indicates how DIF might be assessed for certain constructed-response item types. The description of DIF assessment presented in this paper is applicable to computer-delivered constructed-response items as well as paper and pencil delivered items.

Constructed Response and Differential Item Functioning: A Pragmatic Approach

Neil J. Dorans and Alicia P. Schmitt

From the test practitioner's point of view, constructed-response items transfer the bulk of arduous labor that goes into producing a test item and its score from the test developer to test scorer. Quality multiple-choice items are difficult to produce but easy to score and analyze. Constructed-response items are relatively easy to produce, but difficult to score and complicated to analyze. For multiple-choice items, the psychometrics are well-developed and procedures for performing microscopic dissections of items are well-established. These procedures include techniques for assessing differential item functioning (DIF). Two DIF procedures routinely used at Educational Testing Service are the Mantel-Haenszel procedure (Holland & Thayer, 1988) and the standardization approach (Dorans & Kulick, 1986), both of which are described in detail in Dorans and Holland (in press). Other procedures, based on item response theory (IRT), are described by Thissen, Steinberg and Wainer (in press).

Mislevy, Yamamoto and Anacker (in press) contrast the well-developed body of psychometrics for multiple-choice items with the nascent state of psychometrics for constructed-response items. To the extent that a constructed-response item is unconstrained and examinees are free to produce any response they wish, the test scorer has a difficult and challenging task of extracting information from examinee responses. To date the psychometrics for dealing with this unconstrained response item type have lagged behind the development and administration of these items. Until psychometrics find ways of extracting replicable and valid information from these responses, constructed-response applications will remain the exception in high-volume, "high-stakes" testing applications.

Differential item functioning analysis, which will be defined later, provides secondary psychometrics, usually performed in areas where the primary psychometrics associated with descriptions of item performance, test performance and examinee performance are well-defined, as is the case with multiple-choice items. Given the current state of psychometrics

and cognitive theory for constructed-response testing, we have no alternative but to rely on a class of DIF procedures which is descriptive and can be applied in the absence of cognitive models and related psychometric models, making the class of procedures, in this sense, "model-free."

Constructed-response items move beyond traditional multiple-choice items, for which DIF methodology is well-defined, toward item types involving selection or identification, reordering or rearrangement, substitution or correction, completion, construction, and performance or presentation. Model-free DIF assessment requires both a rule for scoring items and a matching variable on which different subpopulations can be viewed as comparable for purposes of assessing their item performance. This paper focuses primarily on DIF assessment and secondarily on constructed-response. The DIF portions draw heavily upon earlier work, most notably, Dorans and Holland (in press). The standardization and Mantel-Haenszel approaches are described in some detail to give the reader an appreciation of state-of-the-art, model-free DIF assessment and because these procedures can be extended to assess DIF among some constructed-response formats.

The structure of the paper is as follows: DIF is defined and then is contrasted with impact via Simpson's paradox, which demonstrates the importance of matching in DIF studies. The standardization approach is defined as a flexible procedure for describing DIF, while the Mantel-Haenszel (MH) procedure is described as a statistically powerful method for detecting DIF. A common framework from which to view these two related procedures is then presented, from which the essence of model-free DIF is extracted. A general procedure for DIF assessment is outlined, followed by a taxonomy of item responses, and then each option within the taxonomy is evaluated in terms of its amenability to DIF analysis using the general procedure. Next, empirical findings from other studies are discussed in terms of their relevance to constructed-response DIF assessment. Finally, future directions in constructed-response DIF analyses are considered.

Differential Item Functioning

Differential item functioning refers to a psychometric difference in how an item functions for two groups. DIF indicates 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 and differences in group ability.

In the first chapter of the book, *Handbook of Methods for Detecting Test Bias*, Shepard (1982) defines DIF, or what was then called item bias, as psychometric features of the item that can misrepresent the competence of a group. She provides some conceptual definitions of the term offered by other authors, 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 underlies the model-free DIF approaches described in Dorans and Holland (in press) and in this paper.

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 model-based definition underlies the DIF procedures described by Thissen, Steinberg and Wainer (in press).

Thissen (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 for DIF assessment presently employed at Educational Testing Service are the standardization approach (Dorans & Kulick, 1986) and the Mantel-Haenszel approach (Holland & Thayer, 1988). Both procedures compare matched groups and are used primarily with multiple-choice items.

DIF Vs. 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 pervasive 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 score higher than Whites, males score higher than females, and high school juniors and seniors score higher than junior high school students. This difference in performance is called impact.

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 the test on which it appears.

Simpson's Paradox

Simpson's paradox (Simpson, 1951) illustrates why one should compare the comparable, as is done in DIF analyses. Table 1 summarizes the performance of two hypothetical groups, A and B, on an imaginary item.

Insert Table 1 about here

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 show 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%, 12,000 of 24,000, 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 comparable groups 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. The 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 illustrates the importance of comparing the comparable. Both the standardization approach (Dorans & Kulick, 1983, 1986), which has been used on the

Scholastic Aptitude Test (SAT) since 1982, and the Mantel-Haenszel method (Holland & Thayer, 1988), which has been used with most ETS testing programs since 1987, emphasize this principle as well. In practice, both approaches use equal 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. Neither method requires a psychometric or a cognitive model of item or test performance.

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 for DIF. Standardization is described first because of its flexibility and the ease with which it can be generalized to constructed-response DIF assessment.

Standardization: A Flexible Method for Describing DIF

Before the mid-eighties, the Mantel-Haenszel (1959) procedure had not been applied to DIF assessment. Dorans (1982) had reviewed item bias studies that had been conducted on SAT data in the late seventies, and concluded that these studies were flawed because either DIF was confounded with lack of model fit (delta plot approach), or it was contaminated by impact (as a result of "fat matching," the practice of grouping scores into broad categories of roughly comparable ability). A new method was needed, and Dorans and Kulick (1983, 1986) developed the standardization approach.

Standardization's Definition of DIF

An item exhibits DIF when the expected performance on an item differs for matched examinees from different groups. Expected performance 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 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. The focal group is the focus of analysis while the reference group serves as a basis for comparison. At ETS, the current practice is to do analyses in which Whites are the reference group, and Blacks, Hispanics, Asian-Americans, and in some cases, Native Americans, serve as the focal groups, and analyses in which females are the focal group and males are the reference group.

Let $\mathbf{E_f(I|M)}$ define the empirical item-test regression for the focal group *f*, and let $\mathbf{E_r(I|M)}$ 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 $\mathbf{E_f(I|M) = E_r(I|M)}$.

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

$$\mathbf{D_m = E_{fm} - E_{rm}},$$

where $\mathbf{E_{fm}}$ and $\mathbf{E_{rm}}$ are realizations of the item-test regressions at score level m. The $\mathbf{D_m}$ 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, as

measured by total score. Plots of these differences, as well as plots of $E_f(I|M)$ and $E_r(I|M)$, provide visual descriptions of DIF in fine detail. For illustrations of non-parametric item test regressions and differences for a rare actual SAT item which exhibits considerable DIF, see Dorans and Kulick (1986).

Standardization's Primary Item Discrepancy Index

The sheer volume of the SAT item pool precludes sole reliance on plots for DIF assessment. There is a clear need for some numerical index that targets suspect items for close scrutiny, while allowing acceptable items to pass swiftly through the screening process. Standardization has such an index: the standardized p-difference (**STD P-DIF**). This index uses a weighting function supplied by the standardization group to average 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 specific weights for each score level. These are used in weighting each individual D_m before accumulating the weighted differences across score levels to arrive at a summary item-discrepancy index.

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, (i.e., the number of examinees who answer correctly over the 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,

$$\begin{aligned} \text{IMPACT} &= P_f - P_r \\ &= \Sigma_m N_{fm} P_{fm}/\Sigma_m N_{fm} - \Sigma_m N_{rm} P_{rm}/\Sigma_m N_{rm} , \end{aligned}$$

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 the 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. 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 (Dorans & Holland, in press).

STD P-DIF 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.

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 is also quite helpful.

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, in preparation). 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-DIF(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, which means standardization can be used to assess *differential distractor functioning* (Schmitt & Dorans, 1990), *differential speededness* (Dorans, Schmitt, &

Bleistein, 1988; Schmitt, Dorans, Crone, & Maneckshana, 1990), and *differential omission* (Rivera & Schmitt, 1988; Schmitt & Dorans, 1990; Schmitt et al., 1990).

As an example from Schmitt and Dorans (1990), consider the standardized distractor analysis for an SAT antonym item from a disclosed 1984 test form for which the key, distractors and DIF information are provided in Table 2.

Insert Table 2 about here

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%). In addition, standardization indicates to us 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". See Schmitt, Holland and Dorans (in press) for examples in which the standardized distractor analysis corroborates DIF hypothesis for Hispanics.

Mantel-Haenszel: Testing the Constant Odds Ratio Hypothesis

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 DIF. This adaptation is used at Educational Testing Service as the primary DIF detection device. The basic data used by the MH method are in the form of **M** 2-by-2 contingency tables or one large three dimensional **2-by-2-by-M** table.

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. There are also 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 in Table 3.

Insert Table 3 about here

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

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

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.

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,

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

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**,

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

Holland and Thayer (1988) report that the MH approach is the test possessing the most statistical power for detecting departures from the null DIF hypothesis that are consistent with the constant odds ratio hypothesis.

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 on 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 on an item difficulty scale. In general, odds are converted to log odds because the latter is symmetric around zero and easier to interpret.

MH DIF in Item Difficulty Metrics

At ETS, item difficulty estimates in the "delta metric," which has a mean of 13 and a standard deviation of 4, are used by test developers. 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 the metric used by standardization,

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

where,

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

As with the standardization approach, the Mantel-Haenszel procedure does not require a psychometric or a cognitive model of item performance. In this sense, both Mantel-Haenszel and standardization are model-free DIF assessment procedures.

A Common Framework and the Essence of Model-Free DIF

A Common Framework

Up to now, the Mantel-Haenszel method and the standardization method have been described in terms of 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, model-free alternative to item response theory for describing item-ability regressions. The two procedures, however, share a common framework. Dorans (1989) utilized this framework to spell out the similarities and dissimilarities of these two procedures for DIF assessment. Dorans and Holland (in press) demonstrated analytically that for rights-scored tests, 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 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 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 is in the metric for defining DIF. Standardization uses differences in conditional proportions correct $\mathbf{D_m}$, while Mantel-Haenszel uses conditional odds ratios α_m. The second difference is in the choice of weights used to average the $\mathbf{D_m}$ 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. 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. The intuitively appealing focal group frequency distribution, which was employed by Dorans and Kulick (1983) in their original work on the SAT, has continued to be used to describe departures from null DIF. The third difference 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 metric used by standardization has been the p-metric, even with formula-scored tests where an item formula-scored metric would seem superior on logical grounds. In contrast, delta 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 attenuate 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 magnified.

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 these two 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 DIF in slightly different ways:

standardization uses intuitively appealing weighting of conditional differences in proportions correct while the Mantel-Haenszel method uses 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.

The Essence of Model-Free DIF

Although we have just pointed out differences between the statistically powerful Mantel-Haenszel procedure and the flexible standardization approach to DIF assessment, their similarities form the essence of what we are calling empirical DIF assessment, as opposed to the model-based DIF assessment that will be discussed later. First, **both procedures require a well-defined and appropriate matching variable** in order to detect DIF. Inadequate matching variables allow impact to creep back into the results of the DIF assessment. The importance of the matching variable has been discussed often in the DIF assessment literature. The matching variable should measure the same construct as the items being studied for DIF.

Second, both procedures require that **some rule exists for scoring items**. The typical rule is to assign a 1 to a correct answer and a 0 to an incorrect answer though, as Dorans and Holland (in press) demonstrate, the standardization approach is also easy to use with formula-scored multiple-choice items.

Third, both procedures typically use an internal criterion or total score as the matching variable, which implies **the existence of a rule for combining information across items**. An internal criterion is typically employed because the collection of items with which an item is administered often measures a common construct and leads to a single score. For most tests, this combination rule is simply the sum of item scores.

The applicability of existing DIF assessment procedures to constructed-response data hinges on all three of these points, but particularly on the existence of an appropriate matching variable, often obtained by combining information across items, and the existence of a well-defined item scoring rule . If the matching variable exists and the items can be scored

right/wrong, both Mantel-Haenszel and standardization can be used for DIF assessment. For non-binary item scoring, some form of the flexible standardization model can be used for DIF assessment, as can a "successive chops" version of the Mantel-Haenszel method.

Model-Free DIF Assessment for Constructed-Response Items

The importance of the matching variable cannot be overstated in a DIF analysis, especially for constructed-response items which tend to be more time consuming to administer than multiple-choice items. More time per item translates to testing fewer items in a given unit of time, which may imply less reliable internal matching variables. Using an external matching variable may have its own set of problems, as will be discussed. For the purposes of this section, however, a well-defined and appropriate matching variable is assumed.

Nominal Data

Also assumed is that responses to a constructed-response item can be clustered on logical grounds into a limited set of score categories. When the item scoring rule yields only nominal data that cannot be ordered, and when no one category is viewed as "correct" (as in describing alternative response strategies), the standardization procedure can be applied via its standardized differential distractor mode. All response categories are treated in turn "as if" they were the correct response, and proportions choosing each category are computed across focal and reference groups at each score level of the matching variable M. Then, a set of standardization weights can be applied to differences in proportions between focal and reference group members to average differences across levels of the matching variable. This type of analysis, which may have important diagnostic value, can also be used with ordered categories as well.

Successive Binary Chops on Ordered Data

If the scoring rule for a constructed-response item results in an ordered score that ranges from less correct to more correct, and numbers can be attached to each level of this score, then two options are available. Either the binary version of the standardization procedure and the Mantel-Haenszel procedure can be applied to successive binary chops of the data, or a continuous version of the general standardization framework can be applied to these scores.

The successive binary-chop application of existing DIF procedures treats item scores at or above a certain level as correct, while those below that level are incorrect. The binary version of standardization or Mantel-Haenszel is applied to the data at each of several successive chops. Each chop reveals whether DIF is evident at that level of the analyses. A partial application of this notion of successive chops occurs in DIF practice with formula scored tests, where the standardization procedure is applied routinely in binary-chop mode (omits and not reached are treated as incorrect as wrong answers) along with the Mantel-Haenszel method; and for some tests, like the SAT, in formula-scored mode where rights are scored as 1, wrongs are scored as -1/(k-1) and omits and not reached are scored as zero. In fact, the formula-score DIF version of the general standardization model, described in Dorans and Holland (in press), represents an application of the standardization model that illustrates how the method can be used with constructed-response data.

Extended Standardization on Ordered Data

The second approach to ordered data, which provides us with an average DIF value for describing DIF on a constructed-response item, uses the general form of the standardization method. At each matching score level, there exist distributions of constructed-response item scores, I, for both the focal group, (e.g., females), and the reference group, (e.g., males). The expected item scores for each group at each matching score level can be computed by using the frequencies to obtain a weighted average of the score levels. These

expected item scores define the empirical item-test regressions that are the basic building blocks for the standardization approach.

Earlier , we let $E_f(I|M)$ define the empirical item-test regression for the focal group *f*, and let $E_r(I|M)$ define the empirical item-test regression for the reference group *r*. The definition of DIF employed by the standardization approach implies that $E_f(I|M) = E_r(I|M)$. 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.

The standardized p-difference was so-named because the original applications of the standardization methodology defined expected item score in terms of proportion correct at each score level. For the purposes of constructed-response DIF assessment, we let **STD P-DIF** refer more generally to a standardized difference in performance on the item.

For illustrative purposes, suppose responses to a constructed-response item can be clustered into four categories, **A, B, C,** and **D**, and that these four categories receive scores of **9, 8, 7,** and **6**, respectively. These scores could be from essays that were graded on "percent correct" scale. As stated earlier, 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}). Instead of scoring the item 1 if correct and 0 if incorrect or omit as for binary scoring of multiple-choice items, the item is scored 9 if the response is in category A, 8 if the response is from category B, 7 if the response is from category C, and 6 for category D. Under this type of scoring, the expected item performance in the focal group at score level **m** is:

$$E_{fm} = \{A_{fm}*(9) + B_{fm}*(8) + C_{fm}*(7) + D_{fm}*(6)\}/N_{fm} ,$$

where A_{fm}, B_{fm}, C_{fm}, and D_{fm} are counts of the number of focal group members at score level **m** who produced responses in categories **A, B, C,** and **D**, and N_{fm} is the total number of focal groups members at score level **m**. Likewise, for the reference group, we have

$$E_{rm} = \{A_{rm}*(9) + B_{rm}*(8) + C_{rm}*(7) + D_{rm}*(6)\}/N_{rm} .$$

Unlike the **STD P-DIF** for multiple-choice items, this **STD P-DIF** does not range from -1 to +1. Instead, its theoretical range is -3 to +3 as would have been the case had the item been scored 0, 1, 2, and 3.

How Big is Big? A very practical issue that must be addressed with this general approach is the set of flagging rules used to identify too much DIF. Under right/wrong scoring, any difference that exceeds null DIF by 10% in either the positive or negative direction is flagged as large enough to merit careful investigation. With more complicated scoring, (e.g., ordered category scoring as illustrated above), one possibility is to convert all differences to a percent of maximum difference scale, and continue to use the 10% rule (5% for distractor analyses). Another option would be to define the effect size in terms of its ultimate impact on the score that is assigned to examinees. This more sophisticated approach would take into account the number of distinct pieces of information (items) contributing to the reported score, and the importance of the studied item to the combination rule that produces this reported score. More needs to be learned about optimal flagging rules in this context.

The Need for Smoothing. The standardization procedure works well with moderate-to-large data sets, but it runs into trouble with small data sets -- especially when the reference group is small. A standard data analytic strategy for dealing with sparse data is to use some kind of statistical model to smooth away sampling irregularities in the observed data. Ramsay (in preparation) recently developed a kernel-based procedure for smoothing non-parametric

item characteristic curves, such as those used in the standarization method, for binary-scored items. This procedure should readily be adapted to the polytomous case. Ramsay and Holland (in preparation) have also developed a kernel-based procedure for smoothing the log-odds ratio for the Mantel-Haenszel method, which could also readily be adapted to estimated conditional differences in constructed-response item score/matching variable regressions. More theoretical work is needed on these extensions.

The Item Type Continuum, Scoring Rules, and DIF

In this section, we focus on a framework for constructed-response items developed by Bennett, Ward, Rock and LaHart (1990) and evaluate the amenability of these different item types to analysis by our general model-free DIF procedure. Bennett et al. (1990) present a scheme for categorizing item types into seven categories that ranges from the traditional multiple-choice item to the presentation/performance item type. (See also Bennett, tin press). Underlying this taxonomy is what appears to be a continuum ranging from the highly constrained, artificial and easy to score multiple-choice item to the virtually unconstrained, naturalistic and much harder to score presentation/performance item type. Examples of each of these seven item types can be found in Bennett (in press). For DIF purposes, we once again assume that a well-defined and adequate matching variable exists, even though we would question the tenability of this assumption with tenacity in practice.

The Multiple-Choice item type is highly constrained in that examinees are required to choose a single best answer from a limited number of options, usually four or five. For multiple-choice items, the standardization method and the Mantel-Haenszel method can be and have been used repeatedly. In addition, model-based IRT methods (Thissen, Steinberg, & Wainer, in press) have also been used successfully in smaller-scale applications. DIF for multiple-choice items is essentially under control; see Dorans and Holland (in press) for a discussion of some sticky, but small, unsolved problems.

Items in the Selection/Identification class are answered by choosing one or more responses from a stimulus array where the number of choices is large enough to preclude guessing the correct answer. Note that in contrast to the multiple-choice item type, this class of items is less constrained with respect to selection or identification of the correct response. DIF on selection/identification items can be assessed via the model-free approaches of standardization or Mantel-Haenszel or the model-based IRT approaches because these items can easily be scored correct or incorrect. In fact, standard DIF detection techniques probably would work better with this item class than with multiple-choice items because of the near elimination of guessing.

As with selection/identification items, Reordering/Rearrangement responses are chosen from a large stimulus array. The task is to place items in the correct sequence or alternative correct sequence. The elementary probability theory of permutations and combinations tells us that the number of response options grows rapidly when we move from selection/identification to reordering/rearrangement. Hence we have increased the amount of potential diversity or chaos permitted into the response space. Reordering/rearranging items are less likely to be amenable to standard DIF analyses unless possible orders or arrangements can clearly be split into a correct set and everything else. If the sets can be ordered with respect to correctness and different degrees of partial credit awarded to certain sets of responses, then the general model-free standardization method could readily be applied to these items, provided a reasonable clustering of orderings into score categories could be achieved. Even if these categories could not be ordered, the "distractor analysis" could be used to study differential performance on the different clusters.

Substitution/Correction items require that the examinee replace what is presented with a correct alternative. At first glance it appears that this item type has a smaller response space than the reordering/rearrangement item type, but it does not. In fact, it is the first item type in the continuum that has an infinite response space, albeit in practical terms, the number of plausible responses is limited. Substitution/correction items may or may not be amenable to

standard multiple-choice DIF analyses. If the items can easily be scored correct/incorrect, as is most likely the case with correcting grammatical and spelling errors, then the arsenal of DIF techniques that are widely used with multiple-choice items are readily applicable. If partial credit orderings are obtainable or if the categories are only nominal, then the general model-free standardization model can be used.

The Completion item type allows for a slightly greater complexity of responses than the preceding one. Here, the task is to respond correctly to an incomplete, as opposed to incorrect, stimulus. Completion items are probably amenable to standard DIF analyses, especially when there is a clearly defined class of correct answers, as is the case with the mathematical grid-in item type (Braswell & Kupin, in press). If the completion item is carefully crafted, a single of class of equivalent responses should be identifiable. In practice, however, clever examinees may demonstrate that what the test developer thought was a complete set was missing one or two unusual members. A limited set of ordered categories may also be extracted from completion items, in which case the general model-free DIF procedure could be used. Even in the unusual case where only nominal categories could be found, the distractor analysis mode of the general standardization procedure could still be used.

Instead of merely completing a stimulus or correcting one, the examinee presented with a Construction type item has to produce a complete response to a stimulus. The range of possible responses here is very large and the degree of chaos that can swamp any signal in the response space can be imposing. Construction items will be difficult to assess for DIF because they tend to be very time consuming (e.g., a 30-minute essay) and because a large body of literature (Mazzeo, Schmitt, & Bleistein, 1991; Traub & MacRury, 1990) suggests that essays, the most widely-used construction item, do not measure the same construct as multiple-choice items, the most widely-used basis for creating a matching variable. In other words, the matching variable problem which affects all DIF analyses, even those for multiple-choice items, is particularly severe for the construction item types. The best solution to the

problem would be to extract as much relatively independent information from these items via a partial scoring scheme, cumulate the resulting scores across the limited number of items that can be administered, and use this aggregate as the matching variable for a DIF analysis involving the general procedure described earlier. Even here, the limited number of stimuli may preclude using this approach.

Presentation/Performance item types permit the largest amount of freedom on the part of the examinee and, as importantly, allow for the most extraneous noise to enter into the response space. Here, the examinee is required to make a physical presentation or performance delivered under real or simulated conditions in which the object of the assessment is, in some substantial part, the manner of performance and not simply its result. The testing conditions as well as the response options are relatively unconstrained in order to observe a realistic performance. To the extent that the conditions are realistic, they will not be controlled, and comparability across presentations/performances will be hard to achieve. DIF analyses for the presentation/performance item type is probably not possible. In practice, very few of these items will be administered, the testing conditions may be too uncontrolled to permit even consistent scoring across examinees, and the number of examinees tested would probably be too small to permit any reasonable DIF assessment.

In sum, current DIF procedures can be used with multiple-choice items and selection-identification items. These procedures also may work for reordering/rearranging, substitution/correction, and completion. The more general model-free standardization approach would probably be as applicable, if not more applicable, for these three partially constrained item types. DIF analysis for the construction and presentation/performance items types is either very problematic (construction) or virtually impossible (presentation/performance). In the next section we will review results from other studies that support some of the positions we have just expressed.

Empirical Findings Pertinent to DIF Assessment for Constructed-Response Items

Since the late 1960's there has been much interest in comparisons of multiple-choice and constructed-response item formats. Little has been done, however, with respect to comparisons of differential performance by subgroups. In Traub and MacRury's (1990) review, only four studies were referenced. In addition, we located three other studies. All seven studies address the differential performance of males and females only at the total score level. Moreover, the results of these studies are confounded by factors affecting item format comparisons. Traub and MacRury (1990) specify that in order to compare performance, it is essential that item formats be equivalent with respect to trait, scale, instrument, and scoring method. In addition, the selection of an appropriate matching criterion should also be a requirement to avoid confounding DIF with impact.

The purpose of this section is to summarize empirical findings pertinent to constructed-response DIF assessment. First, several studies comparing the performance of men and women on total scores obtained under the two item formats will be briefly described. Second, an illustration of DIF assessment on a constructed-response completion item will be presented. Third, a major constraint that may affect constructed response DIF assessment will be addressed.

Studies Comparing Subgroup Performance on Total Score

Comparison of male and female performance on multiple-choice and constructed-response formats has been reported by Bell and Hay (1987), Bolger (1984), Breland and Griswold (1981), Mazzeo, Schmitt, and Bleistein (1991), Murphy (1980, 1982), Petersen and Livingston (1982), and Schmitt and Crone (1991). All but the last study have focused on total-score test differences.

The studies by Bolger (1984) and Murphy (1980, 1982) were based on comparisons of converted raw scores to percentages of marks attained. They found that males performed

better than females on multiple-choice items than would be expected on the basis of their performance on constructed-response items. Although these researchers attempted to equate the multiple-choice and constructed-response scales by using percentages, this is a questionable scale equivalency method. No instrument equivalency or differences in test performance by ability level were reported.

Bell and Hay (1987) considered item-format ability differences by comparing males and females with either arts or science subject background across an external ability composite. They did not find crossover of the male and female regression lines for the multiple-choice items nor did they find crossover for a particular type of constructed-response item type, comprehension essays. For both of these item types, females outperformed males all along the ability continuum, but the differences were smaller for the multiple-choice items after the raw score differences were converted to percentages in order to permit some degree of comparison. Results for the other constructed-response item type (composition essay) were not easily interpreted because of the interaction between ability and group membership. No special method to attain scale equivalency, other than percentages, was used.

Breland and Griswold (1981) compared male and female performance on several basic skills measures. One of these measures consisted of an English Placement Test with one essay and three multiple-choice sections. Linear regression analyses were computed to compare the prediction equations of males and females using each of the three multiple-choice sections as predictors of the essay portion. They found parallel slopes but different intercepts; for each of the multiple-choice predictors, the expected essay performance of females was higher than the expected essay performance of males across all score levels. Using the reported means and standard deviations on the four sections of the English Placement Test provided by Breland and Griswold (1981), Mazzeo, Schmitt, and Bleistein (1991) calculated standardized differences between males and females and found considerably large differences only for the essay portion of the examination, where females did better than males.

Studies reported by Mazzeo, Schmitt, and Bleistein (1991) on four different examinations of the Advanced Placement Program, and by Petersen and Livingston (1982) on the Admissions Testing Program English Composition Test are all consistent with the previously summarized findings. Relative to males, female examinees perform better on the constructed-response sections than they do on the multiple-choice sections. These findings do not bode well for using scores on multiple-choice sections as a matching variable for DIF analyses on constructed-response items.

An Illustration of DIF on a Completion Type Item

Differential item performance analyses were computed by Schmitt and Crone (1991) on all items of two mathematics examinations consisting of items in both multiple-choice and constructed-response formats. The constructed-response items consisted of a completion type item where students gridded the correct numeric response to regular math computational problems. The total mathematics score, composed of both constructed-response items (grid-ins) as well as multiple-choice items (four-option algebra placement, four-option quantitative comparison and five-option regular math items) was used as an internal matching criterion. DIF analyses indicated that female and Black examinees had differentially lower performance on grid-in items than did comparable groups of males and White examinees. Table 4 presents STD P-DIF summary information for the grid-in items in one of these Mathematics forms. Comparable results were obtained with the other form and with the Mantel-Haenszel delta-difference DIF statistic, as well. Sample sizes for these DIF analyses ranged from 9,943 for White examinees to 641 for Hispanic examinees. There were 7,129 females, 6,088 males, 1,742 Blacks and 728 Asian-American examinees.

Insert Table 4 about here

In Table 4 the 20 grid-in items were categorized into six groups according to their STD P-DIF values. Three of these groups represent positive DIF values (i.e., the focal group

did differentially better than its reference group) and three groups represent negative DIF values (i. e., the focal group did differentially worse than its reference group). |STD P-DIF| values greater than or equal to .10 define the two extreme categories, one for positive DIF and one for negative DIF. The middle categories, one for positive DIF and one for negative DIF, correspond to |STD P-DIF| values between .05 and .10. |STD P-DIF| values between .00 and .05 are categorized in the two least extreme groupings, one for positive DIF and one for negative DIF. Analyses were done to compare the performance on grid-in items between matched White examinees and each of the following focal groups: Blacks, Hispanics, and Asian-Americans; and between matched male and female (focal group) examinees. Although all focal groups demonstrated negative DIF, female and Black examinees had more extreme negative DIF items and larger negative STD P-DIF means across both forms. The percentage of negative items across both forms ranged from 75% to 80% for these two focal groups.

Examination of the multiple-choice items indicated negligible differential performance for females and Blacks on all but one item type, algebra placement. Female and Black examinees demonstrated differentially higher performance on the algebra placement items.

In order to evaluate whether the internal matching criterion was related to the high negative DIF findings for the grid-in item type, DIF analyses for the grid-in items were redone using an external matching criterion which did not include either grid-ins or algebra placement items. The criterion was composed of 60 multiple-choice math items (40 regular math and 20 quantitative comparison). The same six STD P-DIF value groupings used with the internal matching criterion analysis were used to summarize results of this re-analysis. Table 5 presents the classification of grid-in DIF values into the six STD P-DIF groupings for the first form. Results for the second form were comparable.

Insert Table 5 about here

Because the re-analysis using an external matching criterion had to be restricted to those examinees who took the external test, sample sizes for all the groups were considerably

reduced. A minimum sample size of 200 was specified for focal or reference groups. Sample size was insufficient for the Hispanic focal group. For the external criterion analysis, there were 2,717 Whites, 1,992 females, 1,655 males, 527 Blacks and 202 Asian-American examinees.

Results of the DIF analyses using the external all multiple-choice matching criterion did not reduce the magnitude or direction of the DIF found using the internal matching criterion. The external matching criterion seemed to increase the negative DIF found for the Black focal group on the grid-in items of both forms. The STD P-DIF means increased in magnitude from -.03, based on DIF computations with the internal matching criterion, to -.04, based on DIF computations with the external matching criterion. The percentage of negative items across both forms for the Black focal group also increased, from 85% to 95%, as seen in Table 5. Although results for the female/male comparison were not as consistently negative across both forms, they were either basically the same or much worse. These results indicate that there is negative DIF for the grid-in item type using either the internal or external matching criterion.

Dimensionality analyses for both the internal matching test and especially the external matching test indicated that these tests were basically unidimensional in the general population (Lehman & Mazzeo, 1991), which was consistent with earlier, more extensive analyses conducted on other versions of the external matching test (Dorans & Lawrence, 1987). The DIF results, however, question the appropriateness of either the total math test internal or external matching criterion for some subgroups and indicate the possibility that these tests are multidimensional for the subgroups, despite the fact that the tests are unidimensional for the total group. Thus, these total scores might not be an appropriate matching criterion for all item types.

An extreme example of a matching variable which was clearly inappropriate occurred in preliminary analyses done by Mazzeo, Schmitt, and Bleistein (personal communication) where a constructed-response essay section was used as the matching variable to compute DIF

on a multiple-choice section. They found extreme and pervasive negative DIF for the multiple-choice items that had not been evident when the analyses were computed using an internal multiple-choice matching criterion. Obviously, essays and multiple-choice tests did not measure the same construct in the same way or with the same degree of accuracy in this instance.

Matching Criterion

Evaluations of the trait equivalence for multiple-choice and constructed-response items seem to indicate that these two item formats are not equivalent even when the item content and the scoring are maintained across the two item formats. Traub and MacRury (1990) conclude that multiple-choice and constructed-response tests measure different configurations of knowledge and ability. Thus, the total score matching criteria used for constructed-response DIF analyses may, based on these conclusions, have to consist of constructed-response items. If a non-constructed-response matching criterion is used, the comparability of the groups for constructed-response DIF may not be achieved. Because some constructed-response tests, such as those with construction or presentational performance items, consist of a limited number of items, which in turn sample only a very limited domain of knowledge pertinent to the construct being measured, finding an appropriate constructed-response matching criterion may be almost impossible. In such cases, DIF analyses may be impossible to do.

Beyond Description and Detection

Levels of Proficiency and Constructed Response

Mislevy (in press) makes a distinction between **levels of proficiency** and the **architecture of proficiency** when describing psychometric models for educational and psychological test data. In a recent book entitled *A Century of Ability Testing*, Thorndike and Lohman (1990) sketch a history of testing and, as their title implies, most of this testing is of

the levels of proficiency type of testing. Examinees are administered tests, usually multiple-choice but sometimes constructed-response, their responses are recorded and scored, and they are ultimately ordered on a continuum. The model-free DIF procedures that have been described in this paper and elsewhere (Dorans, 1989; Dorans & Holland, in press; Dorans & Kulick, 1986; Holland & Thayer, 1988), as well as those based on IRT models (Thissen, Steinberg, & Wainer, in press) can be used with success on "levels of proficiency" items and tests when the items are binary-scored multiple-choice items. When the level of proficiency items are what Mislevy (in press) calls "something else," (e.g., one of the several constructed-response item types described in this paper), then the general standardization approach can be used, provided that an adequate matching variable, such as level of proficiency, exists along with a well-defined item scoring rule.

Other model-based DIF procedures can undoubtedly be developed from the various IRT models that exist for the non-binary item data case. Thissen and Steinberg's (1986) taxonomy of IRT models is a lodestone for psychometricians interested in developing IRT-based DIF assessment procedures. Thissen and Steinberg make distinctions among: binary models, such as the normal ogive models developed by Lawley (1943), Tucker (1946) and Lord (1952), and the one- and two-parameter logistic models introduced by Rasch (1960) and Birnbaum (1968); difference models, epitomized by Samejima's (1969) "graded-response" model, a pertinent model for constructed-response data; divide-by-total models, such as Master's (1982) "partial credit" model, Andrich's (1978) "rating scale" model, and Bock's (1972) "nominal" model, all of which would seem particularly applicable to "level of proficiency" constructed-response data; left-side added models, epitomized by the three-parameter logistic model (Birnbaum, 1968); and left-side added multiple category models, which modify divide-by-total models to account for guessing, such as is done with Thissen and Steinberg's (1984) "multiple-choice" model.

The binary and left-side added models can be used with binary-scored multiple-choice items. DIF procedures based on these models also exist (Thissen, Steinberg, & Wainer, in

press). The difference, divide-by-total and left-side added models exist for multiple-category responses which are likely to be by-products of well-scored constructed-response items. Our cursory review of the literature uncovered very few DIF applications employing these models, probably because there are few applications involving the type of scoring required for these models, and perhaps more importantly, the absence of readily available and readily usable software that employs these models. Ferrara and Walker-Bartnick (1990) use the "partial credit" model to assess DIF in an essay test, an application involving real data. Thissen, Steinberg and Wainer (in press), unsurprisingly, present an illustrative application of the "multiple-choice" model to assess differential alternative functioning, their expression for differential distractor functioning. More importantly, these authors provide a general IRT likelihood ratio definition of DIF, analogous to the general model-free standardization definition, that holds for all the IRT models described in the Thissen and Steinberg (1986) taxonomy. Thus, in theory, model-based IRT DIF assessment alternatives to the general standardization procedure could be devised for multiple category constructed responses.

The reasons for not using the more elegant model-based DIF assessment procedures for multiple-category responses include some old reasons: complexity, cost, and relative lack of availability of user-friendly software, not to mention lack of understanding and experience. The latter reasons are remediated by proper training and besides, ignorance rarely stands in the way of application, viz., factor analysis in the sixties and seventies and binary IRT in the eighties, and more recently the widespread misuse of the Mantel-Haenszel chi square test as a measure of DIF effect size. The real stumbling blocks will be availability of user-friendly software and cost. As Thissen, Steinberg, and Wainer (in press) point out, the cost of IRT likelihood ratio DIF assessment, as well as other types of IRT-based DIF assessment, is steep compared to the model-free methods of standardization and Mantel-Haenszel.

A more recent issue concerns the desire to avoid confounding model misfit with DIF. The standardization procedure was originally developed as a model-free alternative to IRT-based procedures, which were in the vogue in the early eighties, not because it was less

expensive-- standardization DIF analyses involving hundreds of thousands of examinees (all the data) are cheaper than IRT calibrations involving 1/100th of the data (a spaced random sample of a few thousand)-- but because it was model-free and Dorans and Kulick (1983, 1986) did not want to confound model misfit with DIF. The clearest example of model misfit confounding DIF assessment is the Rasch model, which, like Angoff and Ford's (1973) transformed item difficulty approach, confounds DIF with differences in item acuity or quality. While appropriate IRT models (e.g., the three parameter logistic model), have been shown to be powerful tools in analyzing multiple-choice item data, their applicability to constructed-response data is yet to be demonstrated. Until such time, the prudent course is to use the more descriptive model-free approaches.

Architecture of Proficiency

Mislevy, Yamamoto and Anacker (in press) leave behind the century old world of level of proficiency testing and delve into the relatively unchartered waters of architecture of proficiency (Mislevy, in press). Here the psychometrics are less well-developed. Instead of the simple model of cognitive ability that underlies much of classical test theory and item response theory, namely "the more proficient you are, the better you will do on items and tests of proficiency," these new models attempt to incorporate more complex, maybe more realistic, conceptions of cognition. The authors cite some examples of this new type of psychometric modelling, most of which are theoretical papers or papers involving limited examples. Mislevy and Verhelst (1990) have developed "mixture models" for item responses when different examinees follow different solution strategies or use alternative mental models. This approach involves the identification of classes of examinees who follow distinct solution strategies. Falmagne (1989) and Haertel (1984) employ "binary skills" models which describe competence in terms of the presence or absence of many elements of skill or knowledge. Masters and Mislevy (in press) and Wilson (1989a) use the "partial credit rating scale" model to characterize levels of understanding with respect to its nature as opposed to its

correctness. Wilson (1989b) has described a "Saltus" model to categorize stages of conceptual development by parameterizing the differential patterns of strength and weakness expected as learners progress through successive conceptualizations of a domain. Yamamoto's (1987) "Hybrid" model characterizes an examinee as either belonging to one of a number of classes associated with states of understanding or being placed in a catch-all IRT class. Tatsuoka's "rule space" approach (1983, 1985, 1990) uses a joint distribution of IRT proficiency estimates and indices of lack of fit for individuals to identify systematic patterns of response to particular solution strategies, both correct and incorrect.

The waters of the architecture of proficiency are, as water invariably is, quite fluid. Hence the cognitive and psychometric models are in their early stages of development. Until an island of psychometric understanding emerges from this sea of exploration and innovation, the prudent thing to do is leave the DIF apparatus ashore. As stated earlier, DIF is "secondary" psychometrics that needs a firm psychometric foundation from which to study group similarities and differences. Mislevy (personal communication, October, 1990) views DIF as an unwelcome interaction term with respect to groups connecting observations and inferences, and recommends that we need to have a scoring model or at least a scoring procedure to know whether the interaction term is needed. With levels of proficiency testing, the foundation of scoring models and scoring rules exists. In fact, several alternative solid foundations exist, which partly explains why there are multiple DIF procedures. As the waters of the architecture of proficiency undergo an elemental change and solidify, the time will be ripe for developing DIF procedures based on emerging models. These procedures are likely to be either directly based on the new models or extensions of old procedures, e.g. using a multivariate matching variable with the Mantel-Haenszel and standardization frameworks (Dorans and Holland, in press). The adaptations of Mantel-Haenzel and standardization are likely to be easier to use and less suspect to the side effects of model misfit.

Water has three states: liquid, solid and gaseous. The waters of the architecture of proficiency may appear perilous to many land-locked level of proficiency types. A temptation that must be avoided is to "stay ashore," using binary scoring on constructed responses only because it is easier to defend a simple scoring rule than a complicated one. The cost of constructed response necessitates that we extract as much relatively independent and useful information as we can out of each response. At the very least, we need to use graded responses or ordered multiple category scoring. We may, however, very well need to immerse ourselves in the seas of architecture of proficiency assessment models. Otherwise, constructed-response testing, which is a rather fluid endeavor itself, may vaporize for cost reasons before it ever establishes a firm foothold in the history of testing.

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.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

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

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology, 57,* 212-220.

Bennett, R. E. (in press). On the meanings of constructed-response. In R.E. Bennett, & W.C. Ward (Eds.), *Construction vs. Choice in cognitive measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (RR-90-7). Princeton, NJ: Educational Testing Service

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (Part 5, pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bolger, N. (1984). *Gender differences in academic achievement according to method of measurement.* Paper presented at the annual meeting of the American Psychological Association, Toronto. (ERIC Document ED 255 555).

Braswell, J., & Kupin, J. (in press). Item formats for assessment in Mathematics. In R.E. Bennett, & W.C. Ward (Eds.), *Construction vs. Choice in cognitive measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Breland, H. M., & Griswold, P. A. (1981). *Group comparisons for basic skills measures* (College Board Report No. 81-6). New York, NY: College Entrance Examination Board.

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., & Holland, P. W. (in press). DIF detection and description: Mantel-Haenszel and standardization. In H. Wainer & P. W. Holland (Eds.), *Differential item functioning: Theory and practice* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

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., & 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. (in preparation). *The standardization approach to differential distractor functioning: Assessing differential speededness.*

Falmagne, J-C. (1989). A latent trait model via a stochastic learning theory for a knowledge space. *Psychometrika, 54,* 283-303.

Ferrara, S., & Walker-Bartnick, L. (1990). *Detecting and analyzing differential item functioning in an essay test using the partial credit model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

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

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8,* 333-346.

Holland, P. W. (1985). *On the study of differential item performance without IRT.* Paper presented at the annual meeting of the Military Testing Association, San Diego.

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.

Lawley, D. N. (1943) On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 62-A* (Part I), 74-82.

Lehman, J.D., & Mazzeo, J. (1991, April). *Confirmatory factor analyses of mathematical prototypes.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph No. 7, 17* (4 part 2).

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

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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47, 149-174.*

Masters, G. N., & Mislevy, R. J. (in press). New views of student learning: Implications for educational measurement. In N. Frederickson, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. G. (1991, April). *Do women perform better, relative to men, on constructed-response tests or multiple-choice tests? Evidence from the advanced placement examinations.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Mislevy, R. J. (in press). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett, & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195-215.

Mislevy, R. J., Yamamoto, K., & Anacker, S. (in press). Toward a test theory for assessing student understanding. In N. Frederickson, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Murphy, R. J. L. (1980). Sex differences in GCE examination entry statistics and success rates. *Education Statistics, 6,* 169-178.

Murphy, R. J. L. (1982). Sex differences in objective test performance. *Journal of Educational Psychology, 52,* 213-219.

Petersen, N. S., & Livingston, S. A. (1982). *English Composition Test with Essay: A descriptive study of the relationship between essay and objective scores by ethnic group and sex* (SR-82-96). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institute.

Ramsay, J. O. (in preparation). *Kernal smoothing approaches to nonparametric item characteristic curve estimation.*

Ramsay, J. O., & Holland, P. W. (in preparation). *Smoothing the Mantel-Haenszel estimator to estimate nonconstant odds ratios.*

Ramsey, P. (in press). Sensitivity review process. In H. Wainer & P. W. Holland (Eds.), *Differential item functioning: Theory and practice* (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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34* (4 Part 2).

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., & Crone, C. R. (April, 1991). *Alternative mathematical aptitude item types: DIF issues.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

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,* p. 67-81.

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1990, April). *Differential item omit and speededness patterns on the SAT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (in press). Evaluating hypotheses about differential item functioning. In H. Wainer & P. W. Holland (Eds.), *Differential item functioning: Theory and practice* (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.

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

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10,* 55-73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederickson, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.), *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 H. Wainer & P. W. Holland (Eds.), *Differential item functioning: Theory and practice* (pp. xx-xx). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1984). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thissen, D., & Steinberg, L. (1986). A response model for multiple choice items. *Psychometrika, 49,* 501-519.

Thorndike, R. M., & Lohman, D. F. (1990). *A century of ability testing*. New York, NY: Riverside Publishing Company.

Traub, R. E., & MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. In K. Ingenkamp and R.S. Jager (Eds.), *Tests und trends 8: Jahrbuch der Padagogischen Diagnostik* (pp. 128-159). Weinheim und Basel, Germany: Beltz Verlog.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11, 1-13*.

Wilson, M. R. (1989a). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education, 33,* 125-138.

Wilson, M. R. (1989b). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276-289.

Yamamoto, K. (1987). *A model that combines IRT and latent class models.* Unpublished doctoral dissertation. University of Illinois.

Table 1

Summary of the Performance of
Two Hypothetical Groups on an Imaginary Item

	Group A			Group B		
Ability Level	**Nm**	**Ncm**	**Ncm/Nm**	**Nm**	**Ncm**	**Ncm/Nm**
1	400	40	.10	10000	2000	.20
2	1000	500	.50	10000	6000	.60
3	1000	900	.90	4000	4000	1.0
Weighted Sum	2400	1440	.60	24000	12000	.50

Note. The symbols **Nm, Ncm,** and **Ncm/Nm** 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.

Table 2

The Standardized Distractor Analysis for an SAT Antonym Item from a Disclosed 1984 Test Form

STD P-DIF (Option)			
MA	PR	BLK	
.04	.09	.12	(A) difficult to learn
.00	.00	.00	(B) inferior in quality
.01	.01	.01	(C) providing great support
-.05	-.11	-.16	(D) having little usefulness
.00	.00	.00	(E) feeling great regret

PRACTICAL :

Table 3

2-by-2-by-M Contingency Table for
an Item, Viewed in 2-by-2 Slices

	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}

Table 4

Differential Item Functioning (DIF) Summary
for GRID-IN Items Using an **Internal Criterion**
Math Form A

	Category of DIF Value For All Comparisons					
STD P-DIF Category	CROSS-GROUP Number	CROSS-GROUP % of Items	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISPANIC	WHITE/ ASIAN
			Percent of Items by DIF Category			
DIF ≥ .10	0	0.0	0.0	0.0	0.0	0:0
05≤ DIF < .10	0	0.0	0.0	0.0	0.0	0.0
.00≤ DIF < .05	7	35.0	25.0	15.0	40.0	40.0
-.05< DIF < .00	4	20.0	45.0	60.0	5.0	20.0
-.10< DIF ≤-.05	4	20.0	15.0	15.0	5.0	40.0
DIF ≤-.10	5	25.0	15.0	10.0	0.0	0.0
Total	20	100.00	100.00	100.00	100.00	100.00
Mean	--	---	-0.03	-0.03	-0.01	-0.02
S.D.	--	---	0.05	0.04	0.02	0.04
Maximum	--	---	0.03	0.02	0.05	0.05
Minimum	--	---	-0.12	-0.12	-0.06	-0.09

Table 5

Differential Item Functioning (DIF) Summary
for GRID-IN Items Using an **External Criterion**
Math Form A

STD P-DIF Category	CROSS-GROUP Number	CROSS-GROUP[a] % of Items	**Category of DIF Value For All Comparisons** MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISPANIC[a]	WHITE/ ASIAN
			Percent of Items by DIF Category			
DIF ≥ .10	0	0.0	0.0	0.0	N/A	0.0
.05≤ DIF < .10	5	25.0	5.0	0.0	N/A	20.0
.00≤ DIF < .05	4	20.0	25.0	5.0	N/A	35.0
-.05< DIF < .00	3	15.0	35.0	60.0	N/A	15.0
-.10< DIF ≤-.05	4	20.0	20.0	25.0	N/A	30.0
DIF ≤-.10	4	20.0	15.0	10.0	N/A	0.0
Total	20	100.00	100.00	100.00	N/A	100.00
Mean	--	---	-0.03	-0.04	N/A	0.00
S.D.	--	---	0.05	0.04	N/A	0.06
Maximum	--	---	0.06	0.00	N/A	0.10
Minimum	--	---	-0.12	-0.11	N/A	-0.09

[a]N/A - Insufficient sample size (N < 200) for DIF analysis.