DOCUMENT RESUME

ED 393 914 TM 024 915

AUTHOR Lyu, C. Felicia; And Others
TITLE Smoothed Standardization Assessment of Testlet Level DIF on a Math Free-Response Item Type.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-38
PUB DATE Nov 95
NOTE 35p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Aptitude Tests; *Item Bias; *Mathematics Tests; Multiple Choice Tests; *Nonparametric Statistics; *Test Items; *True Scores
IDENTIFIERS Item Bias Detection; Kernel Method; Scholastic Assessment Tests; *Smoothing Methods; Specification Error; *Standardization; Testlets; Variability

ABSTRACT

A smoothed version of standardization, which merges kernel smoothing with the traditional standardization differential item functioning (DIF) approach, was used to examine DIF for student-produced response (SPR) items on the Scholastic Assessment Test (SAT) I mathematics test at both the item and testlet levels. This nonparametric technique avoids model misspecification problems. It also has fewer sampling errors because of smoothing. Results from the smoothed item-level DIF analysis showed that regular multiple choice items have more variability in DIF values than SPRs. The testlet DIF analysis indicated that results from White examinees may exceed comparable African-American examinees by 14-19 scaled score units on average. Differences between the smoothed standardization and the traditional standardization are small and most likely due to the use of the true score as a matching variable in the smoothed standardization approach. (Contains 2 tables, 4 figures, and 16 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-95-38

ED 393 914

RESEARCH REPORT

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it

Minor changes have been made to improve reproduction quality

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

SMOOTHED STANDARDIZATION ASSESSMENT OF TESTLET LEVEL DIF ON A MATH FREE-RESPONSE ITEM TYPE

C. Felicia Lyu
Neil J. Dorans
James O. Ramsay

Educational Testing Service
Princeton, New Jersey
November 1995

BEST COPY AVAILABLE

TM024915

ERIC

Smoothed Standardization Assessment of Testlet Level DIF on a Math Free-Response Item Type[1]

C. Felicia Lyu[2]
Educational Testing Service

Neil J. Dorans
Educational Testing Service

James O. Ramsay
McGill University

---

[1]This paper was presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1995.

[2]The authors thank Hua-Hua Chang, Ida Lawrence, and Walter Way for their reviews of an earlier version of this paper.

Copyright © 1995 Educational Testing Service All rights reserved

Abstract

A smoothed version of standardization, which merges kernel smoothing with the traditional standardization DIF approach, was used to examine DIF for student produced response (SPR) items on the SAT I Math at both the item and testlet levels. This nonparametric technique avoids model misspecification problems. It also has less sampling errors because of smoothing. Results from the smoothed item-level DIF analysis showed that regular multiple choice items have more variability in DIF values than SPRs. The testlet DIF analysis indicated that White examinees may exceed comparable African American examinees by 14-19 scaled score units on average. Differences between the smoothed standardization and the traditional standardization are small and most likely due to the use of the true score as a matching variable in the smoothed standardization approach.

Introduction

Researchers have developed different techniques for assessing differential item functioning (DIF) during the last decade (Dorans & Potenza, 1994). Most of these techniques focus on DIF for binary-scored items at the item level. We need to investigate differential functioning at the test or testlet or mini-test level for at least two reasons.

First, amplification may occur which could make the DIF more serious than suggested by item analysis. Amplification refers to a situation in which the cumulative effect of DIF in the same direction may be unacceptable at the testlet or test level even though the DIF amount may not be significant for individual items (Nandakumar, 1993).

The second reason for doing a testlet DIF analysis is to match the DIF process with the test construction model. For example, it is reasonable to examine certain items as a set when these items follow a text passage and are interrelated with each other (Wainer, Sireci, & Thissen, 1991). Items of the same item type may have similar properties that warrant simultaneous investigation as well.

Procedures derived from item response theory (IRT) have been used to either assess or interpret differential test functioning (DTF) or testlet DIF. Wainer et al make use of Bock's (1972) polytomous IRT model to compare subgroup differences in discriminations and intercepts against score categories. Shealy and Stout (1993a) interpret DTF from a multidimensional IRT perspective. Their framework distinguishes nuisance constructs from the target construct. The amount of test bias is determined from the expected test score difference between groups at each target ability level. Though Shealy and Stout interpret DTF from a multidimensional-IRT point of view, the relationship between the item score and the target ability is estimated nonparametrically.

In this study, we use the smoothed version of the standardization approach (STND), which we describe later, to examine testlet DIF on free response math items from the new SAT I. Answering student produced response (SPR) items requires a different problem-solving skill because examinees can not extract any information about a possible correct answer from alternatives as they can when answering multiple-choice items. See Braswell (1991) and Fremer (1993) for details about rationales for including the SPR item type in the

new SAT I.

The smoothed STND approach used in this study has joined kernel smoothing with the traditional version of standardization DIF approach. Traditional standardization is flexible and practical. Employing kernel smoothing with traditional standardization has an advantage of removing irregularities in empirical item option curves due to small sample sizes. A smoothed difference in expected SPR subtest scores for different groups is the basic DIF index obtained with the enhanced version of TESTGRAF program (Ramsay, 1995). This study examines the applicability of the smoothed standardization DIF approach to the testlet of SPR items.

Method

STND DIF Procedure

The null-DIF definition for the STND method states that at each level of the matching variable there is no difference in proportions correct between the focal group (the focus of the DIF analysis) and the reference group (the basis for comparison). This can be conceived of as zero difference in expected item score given the matching variable, or as no difference between empirical item test regressions for the focal and reference groups. This approach does not use any parametric function to fit either the empirical item test regressions or the difference between empirical item test regressions of the focal and reference groups. Model misspecification is not a problem, and high collinearity among parameter estimates is not likely to be problematic for STND because as an observed score nonparametric DIF detection procedure the stability of its estimated curves may be reasonable even when the parameter values are unstable due to collinearity.

The general STND approach involves a comparison of two empirical item-test regressions, in which differences in these regressions at each score level are weighted by the relative frequencies of focal group members at that score level. These weighted differences are then summed across score levels to arrive at a measure of DIF. For binary-scored items, the STND index is an average weighted difference in proportions correct (the expected item score under binary scoring) across score levels.

The more general index is STND EY, or standardized expected item score DIF. For the general case, we assume that there is: (1) a matching variable, X, with M levels, $m = 1, 2,\ldots, M$; (2) an ordered item score, Y, with $K+1$ levels, $k = 0, 1, 2,\ldots, K$, and (3) two groups: r (reference) and f (focal).

The general version of STND starts with the computation of conditional expected item scores (Y) at each level of the matching score for both the focal group, $E_{fm}(Y|X)$, and the reference group, $E_{rm}(Y|X)$, via

$$E_{fm}(Y|X) = \sum_k N_{fmk}\, Y_k / N_{fm}, \qquad (1)$$

and

$$E_{rm}(Y|X) = \sum_k N_{rmk}\, Y_k / N_{rm}, \qquad (2)$$

where N_{fmk} is the number of examinees in the focal group at score level m with item score Y_k, and N_{fm} is the total number of examinees in the focal group at score level m of the matching variable, X. The terms N_{rmk} and N_{rm} are parallel reference group frequencies. The item score variable, Y_k, can take on any ordered values , including 0, 1, 2, 3, ... K, or in the case of formula-scoring, $-1/(K-1)$, 0, 1, where K is the number of options.

The next step is to take differences in expected item scores at each level of the matching variable, X,

$$DIF_m = E_{fm}(Y|X) - E_{rm}(Y|X), \qquad (3)$$

and weight these differences by focal group relative frequencies (Dorans & Kulick, 1986), to obtain

$$\text{STND EY-OS} = \sum_m N_{fm} DIF_m / N_f \qquad (4)$$

where OS denotes the observed matching score and N_f is the total number of focal group

examinees.

Smoothing Empirical Item Response Curves

While the STND procedure is relatively free of model misspecification and collinearity problems, it requires sufficient data to directly estimate the item/test regressions, or the ratios of N_{fmk} / N_{fm}. In small samples, this procedure may produce unstable results due to the effects of sampling error. The kernel smoothing procedure, employed in TESTGRAF (Ramsay, 1995), can be used to obtain smoothed versions of the item/test regressions for both focal and reference groups. These smoothed regressions can then be analyzed for DIF, not only on the keyed response, but on all options as well. The kernel smoothing procedure has been used successfully to estimate smoothed versions of empirical item response functions, which are in essence regressions of item score onto some measure of ability (Ramsay, 1991).

Kernel Smoothing.

The TESTGRAF system uses kernel smoothing during the final stages in the process of producing smooth item response functions (Ramsay, 1995). A technical account of this particular application of kernel smoothing can be found in Ramsay (1991). Since applications of this type of nonparametric regression are relatively new in the field of educational measurement, we will present an overview of kernel smoothing that draws heavily from Ramsay (1995). Altman (1992) also offers a readable introduction to the concepts of smoothing and nonparametric regression.

Kernel smoothing is one type of nonparametric regression procedures. It utilizes the notion of **local averaging** to smooth the curve between an independent variable X, such as a total score, and a dependent variable Y, such as an item score. It employs the term, **evaluation point**, to refer to the value of the X, x_q, for which a smoothed estimate of y_q is sought. The collection of these smoothed estimates of y_q define the smoothing function SY. These smoothing functions are obtained by computing an average of those values of Y, corresponding to the values of X which are closest to x_q, the target value or estimation point.

The kernel smoothing employs the rule of local averaging in which a weighted arithmetic mean of the y_m's corresponding to the values of x_m, which are no more than h score units away from x_q, is computed. The term **bandwidth** is attached to the parameter h, which determines the extent to which data are borrowed from adjacent scores levels for smoothing.

Kernel smoothing actually refers to a general class of functions for computing local averages according to different weighting functions. These kernel functions (*KF*) all possess the following properties:

(1)$KF(u)$ is zero or positive for all values of argument u,

(2)$KF(0)$ is the maximum value taken by KF,

(3)$KF(u)$ goes to zero as u deviates more and more in either direction from 0.

The expression for the kernel weighting function, used by Ramsay (1995) in TESTGRAF, is:

$$w_{mq} = KF[(x_m-x_q)/h]/ \sum_m KF[(x_m-x_q)/h], \quad (5)$$

where the denominator, $\sum_m KF[(x_m-x_q)/h]$, ensures that $\sum_m w_{mq}$, equals one. Let $y_m=E_m(Y/X)$, then the kernel smoothing function is defined as

$$SY(x_q) = \sum_m w_{mq} \cdot y_m = \sum_m w_{mq} \cdot E_m(Y/X) \quad (6)$$

Different kernel smoothing functions differ with respect to the mathematical function used for $KF[(x_m-x_q)/h]$. The Gaussian kernel function, $KF(u) = \exp(-u^2/2)$, employs the well-known Gaussian distribution as the weighting function. Most of the weight is assigned to scores close to the evaluation point.

The size of the bandwidth determines the degree of smoothing. Larger values of h yield estimates based on larger sample sizes which reduces sampling variance. At the same

time, these additional Y values are associated with X values other than the point of evaluation (x_q) hence increased bias is introduced. Thus, larger values of h produce more smoothing, more bias, and less sampling variance. In contrast, smaller values of h produce less smoothing, less bias, and retain more sampling variance. The problem of selecting the size of the smoothing parameter is akin to the problem of selecting degree in a polynomial regression model. A tradeoff between bias and reduced variance influences choice of bandwidth.

Fortunately, the tradeoff between bias and sampling variability can be replaced by a concern for the reduction in mean squared error,

$$\text{MSE}[SY(x_q)] = \text{E}[\{SY(x_q) - Y(x_q)\}^2], \qquad (7)$$

where $Y(x_q)$ is the true function relating Y to X. It turns out that MSE is approximately minimized in a wide range of situations by letting h be proportional to $N^{-1/5}$, where N is the total number of examinees (Ramsay, 1995).

Kernel Smoothing of Relationships between Item and Total Score Data

In the basic item analysis problem, the independent variable is some type of summary performance estimate, typically derived from item data, while the dependent variable is the probability of choosing option k for item i. The basic data are y_{ika}, and take on the value 1 if examinee a chooses option k and 0 otherwise. Our goal is to estimate the probability function $P_k(X)$ for each item, and ultimately, from these probability functions, expected item score.

Several steps take place to obtain unsmoothed item response functions before the smoothing steps. First, examinees are sorted and ranked from lowest to highest on the basis of X, which could be a simple number right score, a scale score that is adjusted for differences in test difficulty, or another estimate of proficiency. Next, the a^{th} quantile of the standard normal distribution, z_a, is assigned to the a^{th} examinee in the order of sorted scores. The area of the standard normal density function to the left of z_a is equal to $a/(N+1)$. Then

the same response patterns are sorted by these z_a. Next, the empirical item option functions are computed with respect to the standard normal quantile scores, which serves as input to the kernel smoothing process. A bandwidth, $h = 1.1N^{-1/5}$, is used by default in TESTGRAF.

The smoothed item response curves for the keyed response and each distractor are estimated by smoothing the relationship between the binary indicator variable for each option and the standard normal quantiles. As stated earlier, kernel smoothing is a form of local averaging, in which the weights assigned to the estimate of the curve at z_q are greatest at z_q and taper off to zero as score levels become more distant from z_q. The particular kernel smoothing procedure employed is the Gaussian function with Nardaraya-Watson weights (Ramsay, 1995).

The end products are smoothed empirical item option response curves,

$$SP_k(x_q) = \sum_m w_{kqm} \cdot E_{km}(Y/X) \tag{8}$$

$$SP_k(x_q) = \{\sum_m KF[(z_m - z_q)/h] \cdot y_{km}\} / \sum_m KF[(z_m - z_q)/h], \tag{9}$$

where $y_{km} = E_{km}(Y/X)$.

These smoothed empirical item option curves are averaged via the weighting function used to define expected item score (Dorans, Potenza, & Ramsay, 1994). For rights scoring, the smoothed function for the keyed response is given a weight of one and all other options are given weights of zero. For traditional formula scoring, the keyed response is given a weight of 1, while the sum of the smoothed functions for the distractors is given a weight of $(-1/(K-1))$, where K is the number of options. These two are combined to produce a smoothed estimate of expected item score for the reference group, $SE_r(Y/X)$, and for the focal group, $SE_f(Y/X)$:

$$SE_r(Y/X) = \sum_k SP_{rk} Y_k . \tag{10}$$

$$SE_f(Y/X) = \sum_k SP_{fk} Y_k \,, \qquad (11)$$

where Y_k is the item score variable equal to 0 or 1 for rights scoring, or equal to *-1/(K-1), 0, 1* in the case of formula-scoring, where K is the number of options. Next, one may take differences in expected item scores between the reference group and the focal group at each level of the matching variable and weight these differences by focal group relative frequencies. Then,

$$\text{STND SEY-LV} = \sum_m p_f(sx)\ SDIF_m, \qquad (12)$$

is obtained, where LV denotes the latent matching score, $SDIF_m$ is the difference between the reference group and the focal group in expected item scores at the mth level of the matching variable, and $p_f(sx)$ are the smoothed relative frequencies for the focal group, which will be explained further below. Note that this index parallels the index in equation (4), where the matching variable is the observed score.

Smoothed STND DIF Procedure for Testlets

Provisions have been made in TESTGRAF to compute curves separately by focal and reference groups (Ramsay, 1995). The enhancement to the STND procedure involves the replacement of empirical item response functions with these smoothed item response functions produced by TESTGRAF. An additional option involves the weighting function used to compute the DIF index. The current STND procedure uses the observed frequency distribution of the matching variable in the focal group as the weighting function. The version of smoothed STND-DIF studied here replaces that weighting function with a smoothed estimate of the expected matching variable score in the focal group, which makes it akin to the SIBTEST for bundles procedure developed by Douglas, Stout, and Roussos (1995). Like SIBTEST, the STND approach studied here is what Dorans and Potenza (1994) call a non-parametric latent variable approach for polytomous DIF assessment. The TESTGRAF procedure has also been enhanced to allow for polytomous scoring of items and

the computation of expected item scores, which are used by the expanded standardization technique developed for the assessment of polytomous DIF (Dorans & Schmitt, 1993).

Testlet DIF is simply a special case of that general polytomous scoring model in which a score value is assigned to each possible number correct in the set of items that define the testlet. For example, a perfect score on a set of 10 SPR items earns a value of 10, while 5 correct out of 10 earns a score value of 5, and none correct merits a zero. While more information can be extracted from the pattern of item responses, this simple scoring scheme matches the one used in practice and captures any practical DIF that may occur.

Smoothed Empirical TESTLET Category Curves for Focal and Reference Groups

The first step in the smoothed standardization process for testlets is to use TESTGRAF to produce kernel-smoothed empirical testlet categories for both the focal group and the reference group. The operations described in the preceding section are performed independently for each group, producing estimates of $SP_{fk}(x_q)$ for the focal group and $SP_{rk}(x_q)$ for the reference group. In this case, the options (k) correspond to the different possible number correct score categories from 0 to K for the polytomously scored item, which is actually a testlet containing K items.

Next, the TESTCOMP component of TESTGRAF is used to compare these curves. First, the curves must be expressed in a common metric that is shared across the focal and reference groups. TESTCOMP presently uses expected total or expected formula score, which is obtained by summing the smoothed expected item scores across items, which can be thought of as a double smoothing. Once computed, the best way to examine these curves is visually. When there are many items, as is often the case with large-scale testing programs, numerical indices are needed to help guide the viewing process. Standardization provides such an index,

$$SP\text{-}DTF_k = \sum_m p_f(sx)\ (SP_{fm}(x_q)\text{-}SP_{rm}(x_q)). \qquad (13)$$

Here, $SP\text{-}DTF_k$ is the weighted sum of smoothed proportion differences across m levels of

the matching variable for each testlet category k. The smoothed weighting function $p_f(sx)$ will be defined later in equation (20).

Expected Testlet Scores for Focal and Reference Groups

The smoothed standardization index for the testlet is STND SET, or standardized smoothed expected testlet score DIF. As before, we assume that there is: (1) a matching variable, X, with M levels, $m = 1, 2, \ldots, M$; (2) an ordered polytomous item or testlet score, Y, with $K+1$ levels, $k = 0, 1, 2, \ldots, K$, and (3) two groups: r (reference) and f (focal).

The general version of smoothed STND starts with the computation of expected testlet scores for both the focal group, $SET_{fm}(Y|X)$, and the reference group, $SET_{rm}(Y|X)$, from the smoothed item option curves

$$SET_{fm}(Y|X) = \sum_k SP_{fmk}\, Y_k \qquad (14)$$

and

$$SET_{rm}(Y|X) = \sum_k SP_{rmk}\, Y_k \qquad (15)$$

where SP_{fmk} is the smoothed proportion of examinees in the focal group at score level m with testlet score Y_k, and SP_{rmk} is the parallel reference group smoothed proportion. The testlet score variable, Y_k, can take on any ordered values, including 0, 1, 2, 3, ... K, where K is the perfect testlet score, equal to the number of items in the testlet.

Smoothed Standardized Expected Testlet Score Differences

The current smoothed STND approach may have two weighting options. One possibility is to take differences in expected testlet scores at each level of the matching variable,

$$SDTF_m = SET_{fm}(Y|X) - SET_{rm}(Y|X), \qquad (16)$$

and weight these differences by focal group relative frequencies (Dorans & Kulick, 1986), to obtain

$$\text{STND SET-OS} = N_{fm}SDTF_m/N_f \quad (17)$$

where, N_f is the total number of focal group examinees. The label OS denotes an observed matching score; SET-OS is a variation of the STND procedure in which the item option curves are smoothed but the matching variable is not.

The present version of TESTCOMP employs another weighting option: It uses a smoothed estimate of the expected matching variable score distribution in the focal group as a weighting function for comparing smoothed expected item scores, thus making it a latent variable procedure like smoothed polytomous SIBTEST (Douglas, Stout, & DiBello, 1994). The details of smoothing the weighting function employed in TESTGRAF are as follows.

Smoothing of the Weighting Function

The smoothing of the focal group weighting function uses the quantiles obtained from the early stages in the kernel smoothing process in which examinees are first sorted and ranked from lowest to highest on the basis of X, which could be a simple number right score, a scale score that is adjusted for differences in test difficulty, or another estimate of proficiency. Next, the a^{th} quantile of the standard normal distribution, z_a, is assigned to the a^{th} examinee in the order of sorted scores. Expected total score or formula score is defined as a monotonic transformation of the quantile scores, obtained by summing the individual expected item score curves across all items in the test, which results in a smoothed matching variable, SX. This function can be expressed as

$$sx = g\,(z), \quad (18)$$

and its inverse as

$$z = g^{-1}(sx). \quad (19)$$

Standard scores z are normally distributed, with a density function f(0,1). Hence, the transformation of variable from *Z to SX* is accompanied by a transformation of density function via

$$p_f(sx) = f_t[g^{-1}(sx)]\ (Dg^{-1})\ (sx) \qquad (20)$$

where $(Dg^{-1})\ (sx)$ is the result of evaluating the first derivative of g^{-1} at value sx. This term can be easily approximated numerically. The $p_f(sx)$ are the smoothed relative frequencies for the focal group, which can be used to weight differences in expected item scores at each level of the matching variable. We can obtain a smoothed difference in expected testlet score by placing (16) into

$$\text{STND SET-LV} = \sum_m p_f(sx)\ SDTF_m \qquad (21)$$

where $p_f(sx)$ is the smoothed weighting function for expected matching score, and LV denotes a latent matching variable. Note that one may obtain STND SET-LV by summing up the weighted SP-DTF$_k$ as well.

$$\text{STND SET-LV} = \sum_k SP\text{-}DTF_k\, Y_k \qquad (22)$$

The distinction between (17) and (21) is the use of a smoothed expected total score density estimate in place of the observed relative frequencies. We have presented two versions of smoothed standardized DIF statistics; one uses traditional observed score as the matching variable, STND SET-OS in (17), and the other uses smoothed expected total score as the matching variable, STND SET-LV in (21). The former is a smoothed version of the traditional standardization approach in which the empirical item options curves are smoothed via kernel smoothing, while the latter also replaces the weighting function, focal group relative frequencies on the matching variable, with a smoothed expected matching score distribution in the focal group. The latter approach is used in this study for assessing testlet

DIF for free response items.

Analysis and Results

Data

Two SAT I-mathematical tests were evaluated for differential functioning between African American and White examinees on ten SPR items for each test. Random samples were drawn from the total population. For Test 1, the numbers of White examinees and African American examinees are 16,949 and 2,399 respectively. For Test 2, the numbers of White examinees and African American examinees are 9,942 and 4,836. In addition to the ten SPR items, each mathematical test contains 35 Regular Multiple Choice items (five choices) and 15 Quantitative Comparison (QC) items (four choices).

STND DIF and Testlet DIF Analysis

First, the traditional STND DIF statistic (STND EY-OS) is computed for each item. This index is the expected item score difference matched on the observed score. Correspondingly, the smoothed STND SEY-LV statistic is also computed for comparison. These two approaches have two differences: The STND SEY-LV approach smooths the nonparametric item/test regressions, while the traditional procedure does not. The STND-SEY-LV approach uses the expected formula score as the matching criterion, which sums up 60 expected item scores, and smooths the focal group weighting function. It is a latent variable approach. The traditional STND procedure uses the observed formula score instead, making it an observed score approach. For STND SET-LV, the smoothed probability curves for each testlet category and the expected testlet score curves are produced via TESTCOMP.

The TESTGRAF Procedures for Testlets

The TESTGRAF program estimates the testlet DIF in this study by going through the following steps separately for the focal and reference groups:

Step 1: A proficiency measure, the total test formula score, was obtained for each examinee. Examinees were sorted based on the formula score value. The SPR testlet score (0 to 10) was input as a polytomously scored item in the program, as well as other multiple choice items.

Step 2: Each examinee was assigned a quantile value of the standard normal distribution by the ranking order of the formula score. For each testlet score level or category (partial credit in the polytomously scored item), an indicator 0/1 was computed to indicate whether this examinee's SPR testlet score was equal to this category or not.

Step 3: A regression was estimated by smoothing the relationship between the 0/1 indicator and the standard normal quantiles for each testlet category. The end product was the smoothed probability curve for each testlet category. The expected testlet score was computed by assigning each testlet category its score value (0-10) and summing up across all testlet categories.

Step 4: The matching variable was computed by summing up all of the expected item scores and the expected testlet score.

Step 5: Smoothed weighted differences were computed between two comparison curves for testlet categories (SP-DTF), for expected item scores (SEY-LV), and for expected testlet scores (SET-LV), using the smoothed expected total score frequency distribution for the focal group. The index corresponding to SP-DTF at the item level was computed also. That is denoted by SP-DIF, the smoothed weighted difference between empirical item option curves for the reference and focal groups.

Test 1 Analysis

Item-level DIF statistics for ten SPR items are reported in Table 1. Both traditional and smoothed STND produced similar results. Five items have slightly larger STND EY-OS than smoothed STND SEY-LV in absolute values. The traditional approach seems to exhibit slightly more DIF. No SPR items exhibit sizeable DIF (an absolute value greater than .10). However, nine of the ten items are definitely negative. This consistent negative DIF values suggest an examination of amplification effect. The means for other item types are positive and closer to zero (.01, and -.00 for Regular 5-choice items;.02, and .01 for QC items) than

the SPR means (-.02, and -.03).

Figure 1 presents box plots for a fuller description of the DIF distributions. The 75th and 25th percentiles of DIF values are portrayed by the top and bottom of the rectangle, and the median is portrayed by a horizontal line segment within the rectangle. A solid line extends from the top end of the box to the largest observation that is less than or equal to the upper quartile plus 1.5 times of interquartile range. The same holds for the bottom end of the box but in a different direction. Any item which has DIF falling outside the range is plotted as a circle.

Unlike the other item types, SPR items have a central interquartile that falls completely below zero. The central bulk of SPR DIF are compressed with a short length of the box, and tails do not extend as far as the other two item types. The smoothed and traditional indices have similar distributions for SPRs except that the central DIF data for the traditional indices are shifted slightly downward. For multiple choice items, both Regular 5-choice and QC items have medians above zero. Longer boxes and stretched tails indicate that multiple choice items have more variation in DIF values, especially QC items. Generally, traditional STND EY-OS indices exhibit smaller absolute DIF than smoothed STND SEY-LV when DIF is positive and larger absolute DIF when DIF is negative.

Figure 2 presents expected testlet score curves and smoothed probability curves for different testlet categories in five panels. The solid lines are curves for African American examinees, while the dotted lines are curves for White examinees. With an exception of the middle one, the four panels contain smoothed estimated probability curves for SPR testlet scores equal to 0, 2, 5, and 9. Above each of the four panels is the numerical index of STND SP-DTF: smoothed proportion differences for the testlet category. The probability of getting all SPR items incorrect (SPR testlet score = 0) is higher for African American examinees than for White examinees at low ability levels. In contrast, the probability of getting 9 SPR items correct is higher for White examinees than for African American examinees at high ability levels. The two curves for 9 items correct are noticeably different between expected formula score 30 and 50. But data are sparse for African Americans in this area and thus the computed STND SP-DTF is equal to 0. For SPR testlet scores equal to 2 and 5, the curves cross in the middle. The cross-over phenomenon cancels out

differential functioning and we have STND SP-DTF close to 0 again.

The middle panel displays smoothed curves for the expected SPR testlet score. Above the plot is the testlet DIF index STND SET-LV. African American examinees have lower expected SPR test scores than comparable White examinees through all the expected formula score range. The curve for African Americans stops around expected score 50 because there is little data above 50 for African Americans. The testlet DIF index STND SET-LV is -.23. The corresponding traditional STND testlet DIF is -.27, the sum of the ten item DIF values. To make this difference more interpretable, we transformed the unsmoothed testlet DIF value into the 200-800 scale and obtained about 19 scaled score units. This number was obtained by dividing the unsmoothed testlet DIF value by the focal group standard deviation for the SPR testlet score, then adjusting this value by the ratio of total test reliability and SPR testlet reliability for African American examinees, and then multiplying by the scaled score standard deviation of 110.

Test 2 Analysis

Similar to Test 1, eight of ten SPR items in Test 2 have negative DIF though they are small in magnitude (see Table 2). Means and standard deviations for SPR items are the same for both traditional and smoothed DIF indices. The QC items have higher mean DIF with the smoothed STND SEY-LV approach (=.03). Similar to Figure 1, Figure 3 shows that SPR has an interquartile range below zero, lower than other item types. The traditional STND EY-OS interquartile range and median for SPRs are lower than those for smoothed STND SEY-LV. Unlike Test 1 where the minimums were similar, the minimum SPR DIF from smoothed STND is lower than from traditional STND in Test 2. For the other two item types, the DIF distributions are more compressed than they were with Test 1.

Figure 4 presents five plots for Test 2. These plots are similar to the five plots in Figure 2 for Test 1 except that the two curves in the middle panel cross a little bit at expected formula score 52. The STND SET-LV value is -.23 for Test 2, same as Test 1. The corresponding traditional STND testlet DIF is -.25, close to the smoothed version. The traditional DIF statistic transforms to about 14 scaled score units.

Discussion

This study applied the smoothed standardization approach for free response items in the SAT I mathematical tests to assess item DIF and testlet DIF between African American examinees and White examinees. The DIF procedures used in this study have smoothed the item-test regression, as well as the focal group weighting function of the matching variable. The matching variable is the sum of expected item scores, a latent variable rather than an observed variable used in the traditional STND procedure. The smoothed STND procedure produced larger item-level DIF when DIF is positive and smaller DIF when DIF is negative than the traditional STND approach. This difference between traditional STND DIF and smoothed STND DIF is due to use of a true score in place of an observed score as a matching variable rather than smoothing.

Multiple choice items have more variability in DIF than SPR items. This is because the item-level DIF index presented in this study is based on formula scoring: 5-choice items take values -1/4, 0, 1; 4-choice items take values -1/3, 0, 1, and SPRs are 0 and 1 scored. These differences in item scoring rules lead to differences in DIF variability.

The fact that QC items display more positive DIF may be related to differential subgroup guessing behaviors associated with fewer options in QC items. More empirical analyses of differential guessing or differential omission are needed.

The relatively large proportions of negative item level DIF for SPR items has suggested an investigation of amplification effect at the testlet level. The same testlet DIF behavior has been observed for the two tests: African American examinees performed lower on SPR items than comparable White examinees through all of the score range.

The TESTGRAF program produced proportion difference curves for each expected SPR testlet category. The existing cross-over in curves for the middle SPR score levels shrink the STND SP-DTF indices due to cancellation. Though the STND SP-DTF index may not describe the curve adequately, the conjunction of this index with the assigned testlet score values does describe how much that category contributes to overall testlet DIF.

A general formulation of smoothed standardization was presented for both item-level and testlet-level differential functioning. The DIF amount can be assessed by TESTCOMP

for both binary and polytomous data. DIF indices for the case of multiple focal groups can also be developed. More simulation work is need to evaluate standard error formulas, and observed score vs. expected score matching when both are used in the smoothed regression situation.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*, 175-185.

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

Braswell, J. (1991). *Rationales for current and proposed SAT mathematics item types.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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., & Potenza, M. P. (1994). *Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning.* (RR-94-49). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Potenza, M. P., & Ramsay, J. O. (1994). *Smoothed Standardization: A Small Sample DIF Procedure.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, 1994.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Douglas, J., Stout, W., & DiBello, L. (1994). *Smoothed SIBTEST: A new estimator and hypothesis test of DIF.* Unpublished manuscript.

Douglas, J., Roussos, L., & Stout, W. (1995). *Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF.* Unpublished manuscript.

Fremer, J. J. (1993). *What should determine the content of the SAT?* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test

for DIF. *Journal of Educational Measurement, 30,* 293-312.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve problems. *Psychometrika, 56,* 611-630.

Ramsay, J. O. (1995). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data.*

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

Shealy, R.T., & Stout, W. F. (1993b). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Table 1

Smoothed and Traditional Standardization

DIF Statistics for SPR Items in Test 1

	Smoothed STND-SEY-LV	Traditional STND-EY OS
SPR ITEMS		
1	-.05	-.05
2	.04	.04
3	-.05	-.06
4	-.04	-.05
5	-.03	-.03
6	-.03	-.04
7	-.03	-.03
8	-.03	-.05
9	-.01	-.01
10	-.00	-.01
Mean	-.02	-.03
S.D.	.03	.03
Regular 5-Choice Items (35 items)		
Mean	.01	-.00
S.D.	.04	.04
QC Items (15 items)		
Mean	.02	.01
S.D.	.04	.04

Table 2

Smoothed and Traditional Standardization

DIF Statistics for SPR Items in Test 2

	Smoothed STND-SEY-LV	Traditional STND-EY-OS
SPR ITEMS		
1	-.07	-.05
2	-.05	-.06
3	-.00	-.00
4	.03	.04
5	-.02	-.03
6	-.03	-.05
7	-.04	-.05
8	-.02	-.04
9	-.01	-.02
10	.00	.01
Mean	-.02	-.02
S.D.	.03	.03
Regular 5-Choice Items (35 items)		
Mean	.01	-.00
S.D.	.03	.03
QC Items (15 items)		
Mean	.03	.01
S.D.	.03	.03

Figure 1

Distributions of Smoothed and Traditional STND DIF Statistics by Item Type for Test 1

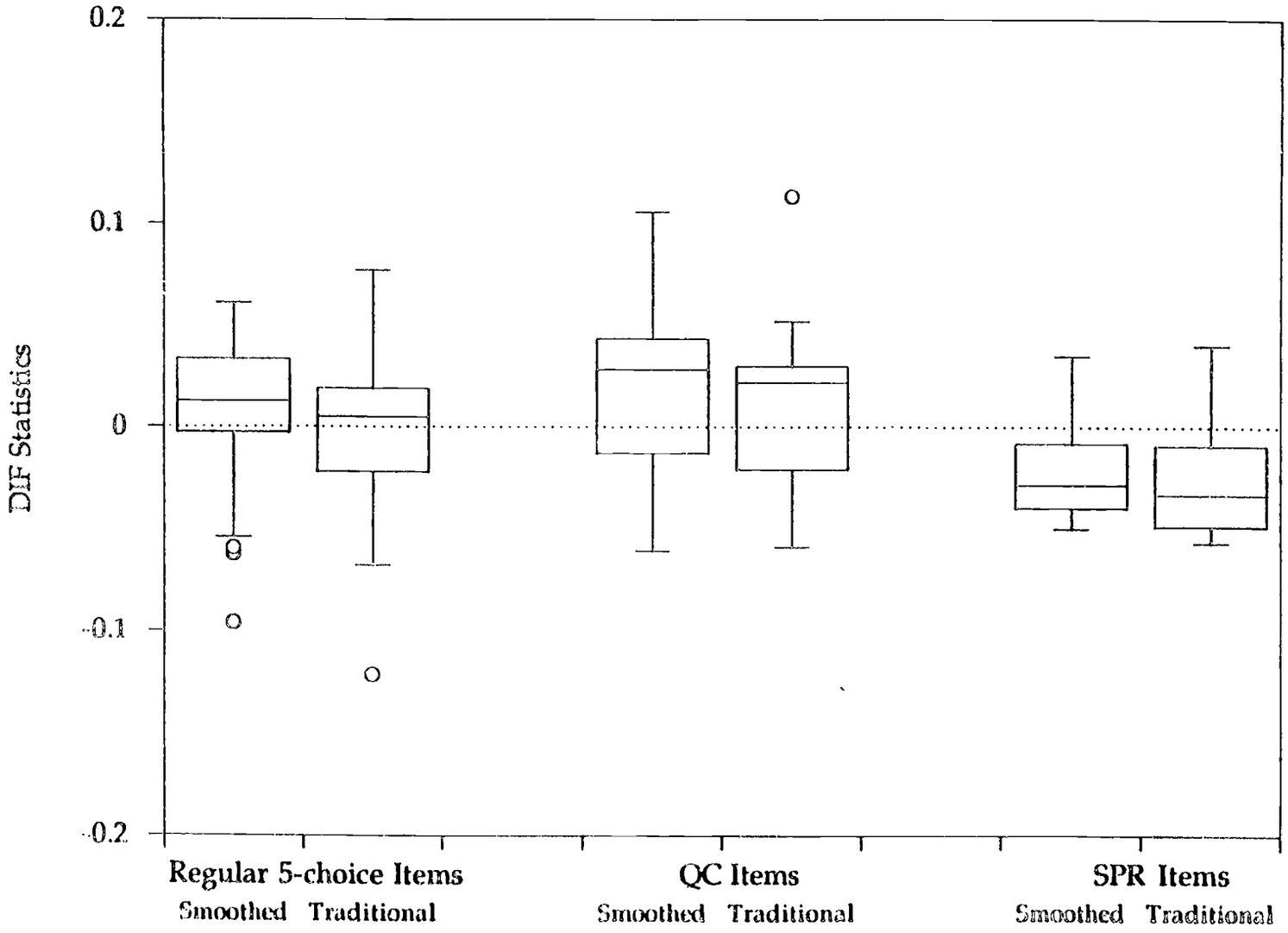
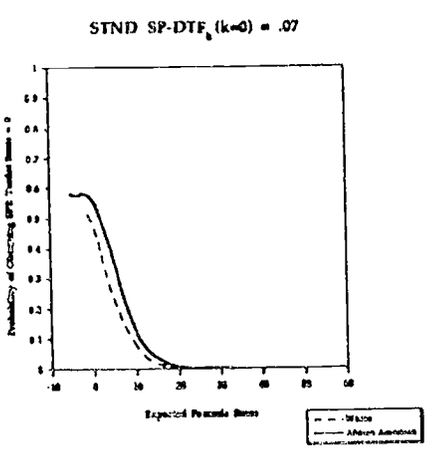
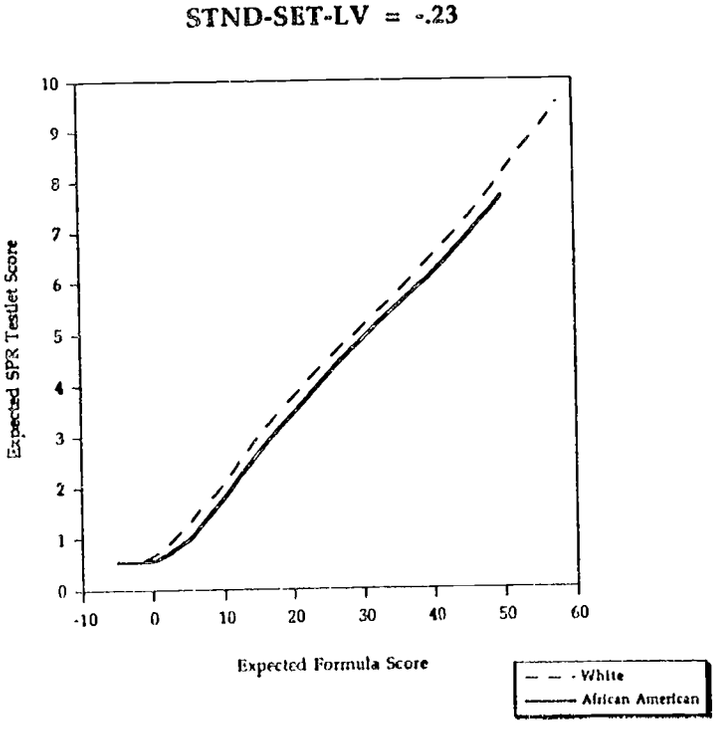
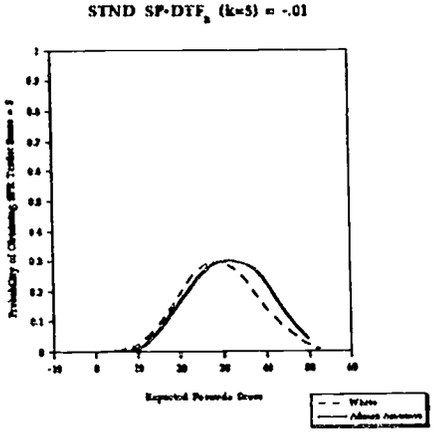


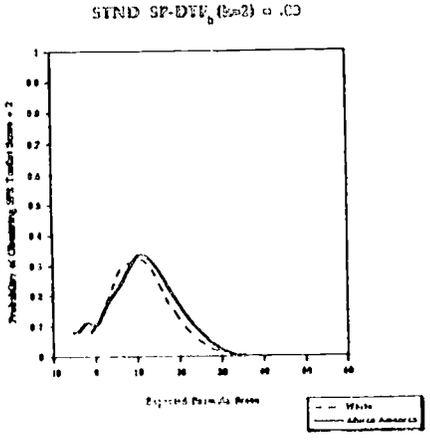
Figure 2

Expected SPR Testlet Scores and Probability Estimates for Testlet Categories in Test 1 with Smoothed STND Approach









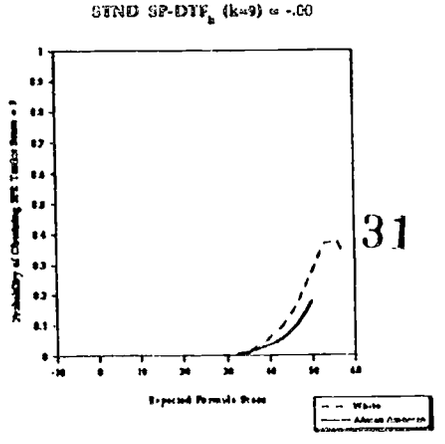


Figure 3

Distributions of Smoothed and Traditional STND DIF Statistics by Item Type for Test 2

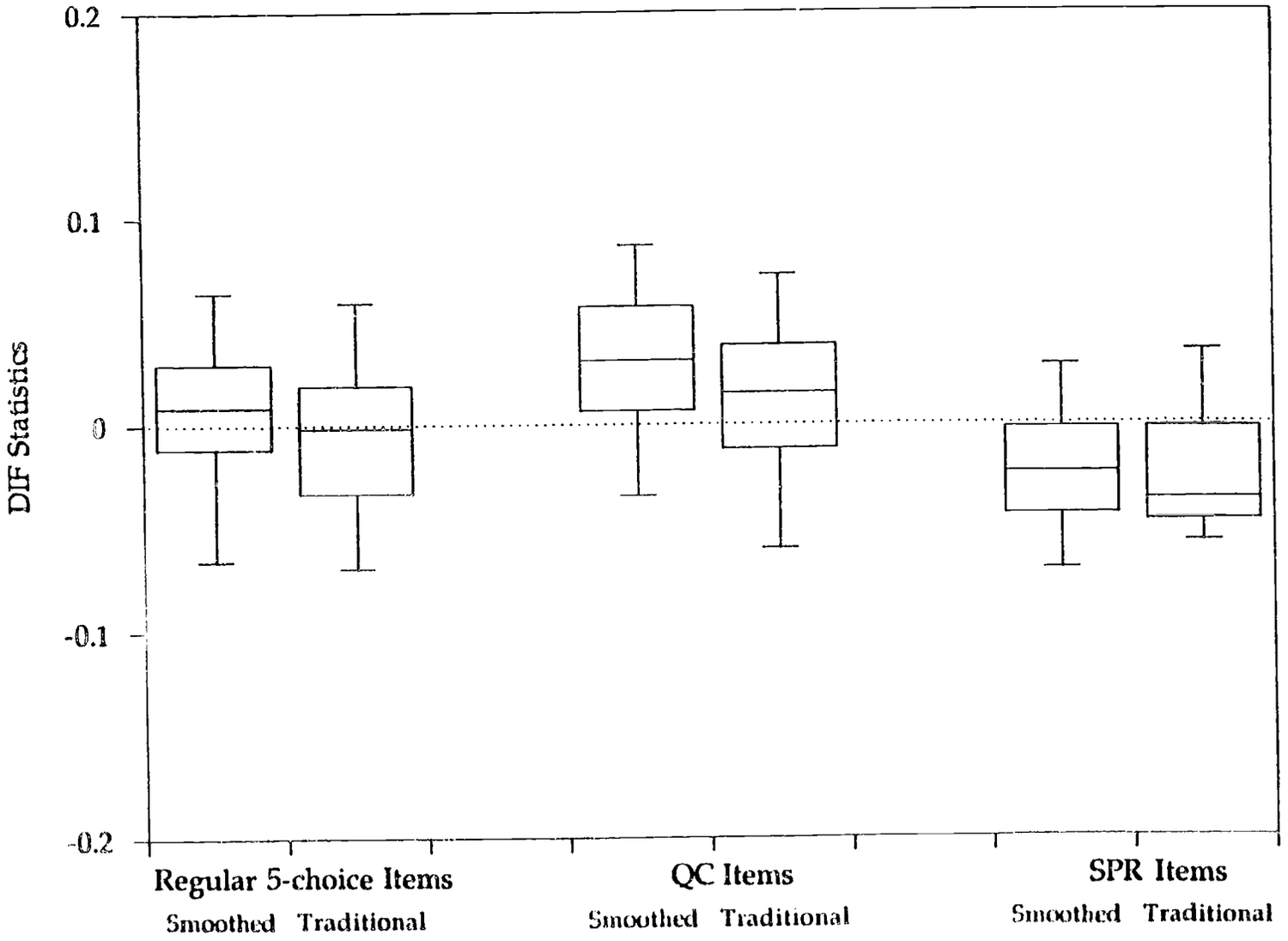


Figure 4

Expected SPR Testlet Scores and Probability Estimates for Testlet Categories in Test 2 with Smoothed STND Approach