Listening. Learning. Leading.®

Research Report
ETS RR-13-33

# Constructed-Response DIF Evaluations for Mixed-Format Tests

Tim Moses

Jinghua Liu

Adele Tan

Weiling Deng

Neil J. Dorans

December 2013

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

# Constructed-Response DIF Evaluations for Mixed-Format Tests

Tim Moses, Jinghua Liu, Adele Tan, Weiling Deng, and Neil J. Dorans
Educational Testing Service, Princeton, New Jersey

December 2013

**Action Editor:** Marna Golub-Smith

**Reviewers:** Rui Gao and Rebecca Zwick



**Abstract**

In this study, differential item functioning (DIF) methods utilizing 14 different matching variables were applied to assess DIF in the constructed-response (CR) items from 6 forms of 3 mixed-format tests. Results suggested that the methods might produce distinct patterns of DIF results for different tests and testing programs, in that the DIF methods' results might be similar for tests with multiple-choice (MC) and CR scores that are similar in their measurement characteristics but would exhibit larger variations for tests with MC and CR scores having more distinct measurement characteristics. Impact measures of the MC and CR scores appeared to be a useful basis for indicating the scores' measurement similarity, for predicting the variations of DIF results from using these scores as matching variables, and possibly for indicating the most appropriate DIF method and matching variable for a particular test. The results are described in terms of their implications for research and practice.

Key words: constructed response, differential item functioning, DIF, mixed-format tests

Evaluations of differential item functioning (DIF) in constructed-response (CR) items have been fairly limited in ETS testing programs in spite of considerable research attention (Chang, Mazzeo, & Roussos, 1996; Dorans & Schmitt, 1993; Kim, Cohen, Alagoz, & Kim, 2007; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Penfield, 2007; Penfield & Algina, 2006; Zwick, Donoghue, & Grima, 1993; Zwick, Thayer, & Mazzeo, 1997). A statement made in a study conducted more than 15 years ago is an accurate description of current CR DIF practice: "At present, ETS has no official policy for screening polytomous items for DIF" (Zwick et al., 1997, p. 1). A survey of the statistical coordinators and managers for ETS testing programs indicated that out of 26 testing programs that administer CR items, only eight routinely evaluate them for DIF. For testing programs that do not conduct routine CR DIF evaluations, the reasons included lack of clarity about what matching variable to use, lack of clarity about the flagging rules, and small sample sizes.

The ambiguities about CR DIF analyses are especially apparent for mixed-format tests, which typically contain a relatively large number of multiple choice (MC) items and a smaller number of CR items. Mixed-format tests present a choice of the matching variable to use to match the focal and reference groups when evaluating these groups' CR item scores for DIF. The CR items being evaluated for DIF are typically assumed to be more similar in their measurement to other CR items than to the MC items, implying that a total CR score (i.e., CR that includes *Y*, the studied item being assessed for DIF) would be a more appropriate matching variable than an MC score. Other issues present additional complications, such as, CR tests are often much shorter and less reliable than MC tests, even when a CR item being evaluated for DIF is included in the CR score being used as the DIF matching variable. Relatively low CR reliability can actually result in the MC scores being more highly correlated with *Y* than the CR scores, implying, for some mixed-format tests, that the MC score may be a potentially better DIF matching variable than the CR score.

In CR DIF research and practice, the choice of matching variable is often either a total test score (MC + CR) or the total test score that excludes the studied item (MC + CR - *Y*). The use of the total test score has been recommended in research on DIF methods that use matching variables in their observed form. The use of total test scores in their observed form is justified in terms of obtaining the most accurate DIF results when *Y* follows a partial-credit model and has no DIF (i.e., better Type I error maintenance; Penfield, 2007; Penfield & Algina, 2006; Zwick et

al., 1993; Zwick et al., 1997). Other DIF methods such as the PolySIB (simultaneous item bias) test use an estimated true-score version of the matching variable that excludes the studied item, *T*(MC + CR - *Y*). The PolySIB's true-score estimation approach addresses the unreliability of the matching variable and the accuracy problems created by matching variable unreliability when there are focal versus reference group differences on that matching variable (Chang et al., 1996).

Based on the previously described survey of statistical coordinators at ETS, the possible statistical characteristics of MC and CR scores in mixed-format tests, and the suggestions of DIF method implementations from prior research, it appears that valid arguments could be made for using multiple DIF methods and matching variables to assess CR DIF in mixed-format tests. In this study, several CR DIF methods are applied and compared to evaluate CR DIF in six forms of three mixed-format tests. The goal is to show how CR DIF evaluations may be usefully implemented as comparisons of variations of methods currently used in ETS testing programs (i.e., the standardized estimated DIF [E-DIF] and PolySIB procedures; Chang et al., 1996; Dorans & Schmitt, 1993). Two questions are of interest:

1. How are the results from various DIF methods and matching variables likely to differ in actual ETS testing data?
2. What data characteristics are most useful for interpreting the results from different CR DIF methods?

This study's comparisons and analyses are extensions of practice and prior research, which produce recommendations for improving future practice and for broadening CR DIF research.

## Method

This study's issues are addressed by developing and applying 14 DIF methods to evaluate CR items for gender DIF in six forms of three mixed-format tests. The 14 DIF methods include seven implementations of the standardized E-DIF and PolySIB DIF methods, because an initial survey of ETS testing programs indicated that the eight programs that routinely assess CR DIF use these methods. The seven implementations of the standardized E-DIF and PolySIB tests are based on seven matching variables: the total CR score, the CR score excluding the studied item (CR - *Y*), the MC score, the MC + CR score, the MC + CR - *Y* score, the bivariate (MC, CR) score combination, and the bivariate (MC, CR - *Y*) score combination. The descriptions and

notations of the 14 considered DIF methods are summarized in Table 1 and described in more detail in the following section. Two forms of the *SAT*® Math test and two *Praxis*™ test titles were considered in this study.

**Table 1**

***Definitions and Notations of the 14 DIF Matching Variables***

| Matching variable | As used in PolySIB | As used in the standardized E-DIF |
|---|---|---|
| Score on all CR items except studied item *Y* | *T*(CR - *Y*) | CR - *Y* |
| Score on all CR items including studied item *Y* | *T*(CR) | CR |
| Score on all CR and MC items except studied item *Y* | *T*(MC + CR - *Y*) | MC + CR - *Y* |
| Score on all CR and MC items including studied item *Y* | *T*(MC + CR) | MC + CR |
| Bivariate score combination on MC items and CR items except studied item *Y* | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* |
| Bivariate score combination on MC items and CR items including studied item *Y* | *T*(MC), *T*(CR) | MC, CR |
| Score on all MC items | *T*(MC) | MC |

*Note.* PolySIB = simultaneous item bias; E-DIF = expected DIF; CR = constructed response; MC = multiple choice.

**Constructed-Response DIF Methods**

All of the CR DIF methods considered in this study can be summarized in terms of an average difference in expected and conditional scores of the studied item (*Y*) for reference (*G* = *R*) and focal (*G* = *F*) groups matched across the *j* = 1 to *J* possible score values of a matching variable,

$$\sum_j \left( \frac{n_{j,F}}{N_F} \right) \left[ E(Y \mid Matching_j, F) - E(Y \mid Matching_j, R) \right], \quad (1)$$

where $n_{j,F}$ and $N_F$ denote the focal group's conditional and overall sample sizes. Equation 1 can be used to express several considered CR DIF methods. CR DIF methods based on standardized E-DIF (Dorans & Schmitt, 1993) use expected and conditional *Y* scores computed as conditional means,

$$E(Y \mid Matching_{\,j}, G) = \mu_{Y|j,G}, \tag{2}$$

where $\mu_{Y|j,G}$ denotes the conditional mean of *Y* for the *j*th score of the matching variable in group *G*. The five matching variables used in Equation 2 are the observed CR, CR - *Y*, MC + CR, MC + CR - *Y*, and MC scores.

CR DIF methods based on PolySIB (Chang et al., 1996) use expected and conditional *Y* scores that are adjusted and interpreted as conditioned on $T(Matching_j)$, the reference and focal groups' estimated true score for the matching variable's *j*th observed score,

$$E[Y \mid T_G(Matching_{\,j})] = \mu_{Y|j,G} + \left[\frac{\mu_{Y|j+1,G} - \mu_{Y|j-1,G}}{T_G(Matching_{\,j+1}) - T_G(Matching_{\,j-1})}\right]\left[T(Matching_j) - T_G(Matching_j)\right], \tag{3}$$

where $T_G(Matching_j) = \mu_{Matching|G} + rel(Matching_{.,G})(Matching_j - \mu_{Matching|G})$, $rel(Matching_{.,G})$ denotes the alpha reliability or internal consistency of the matching variable in group *G* (Kelley, 1923; Shealy & Stout, 1993), and where $T(Matching_j) = \frac{T_R(Matching_j) + T_F(Matching_j)}{2}$. The five matching variables used in Equation 3 are the estimated true scores, *T*(CR), *T*(CR - *Y*), *T*(MC + CR), *T*(MC + CR - *Y*), and *T*(MC).

Prior to computing gender DIF estimates based on Equations 1–3, the male (reference) and female (focal) test data were smoothed using loglinear models (Holland & Thayer, 2000). The use of smoothed frequency data resulted in more stable CR DIF estimates and increased estimation accuracy (Moses, Miao, & Dorans, 2010) and also made it unnecessary to use some data exclusion practices recommended for SIBTEST methods like PolySIB (e.g., data would not be excluded from the SIBTEST calculations when the reference and focal groups' sample sizes were less than two at any score of the matching variable; Shealy & Stout, 1993).

**Bivariate CR DIF Matching Variables**

In addition to using Equations 1–3 to evaluate CR DIF based on the 10 previously described matching variables, the CR DIF methods are also extended to include four additional bivariate matching variables based on the joint distributions of the MC and CR or CR - *Y* scores. The standardized E-DIF versions of Equation 2 based on bivariate matching of these (MC, CR - *Y*) and (MC, CR) distributions are

$$E[Y \mid Matching1_j, Matching2_k, G] = \mu_{Y|j,k,G}. \tag{4}$$

The PolySIB versions of Equation 3 based on bivariate matching of the [*T*(MC), *T*(CR - *Y*)] and [*T*(MC) *T*(CR)] distributions are

$$\begin{aligned} &E[Y \mid T_G(Matching1_j), T_G(Matching2_k)] = \\ &\mu_{Y|j,k,G} + \left[\frac{\mu_{Y|j+1,k,G} - \mu_{Y|j-1,k,G}}{T_G(Matching1_{j+1}) - T_G(Matching1_{j-1})}\right]\left[T(Matching1_j) - T_G(Matching1_j)\right] + \\ &\left[\frac{\mu_{Y|j,k+1,G} - \mu_{Y|j,k-1,G}}{T_G(Matching2_{k+1}) - T_G(Matching2_{k-1})}\right]\left[T(Matching2_k) - T_G(Matching2_k)\right]. \end{aligned} \tag{5}$$

Equations 4 and 5 can both be described as nonlinear regressions where *Y*'s conditional means are related to the joint score combinations of the two matching variables. Equation 5 is especially analogous to multiple linear regression models (Pedhazur, 1997) where *Y*'s conditional means are functions of partial slopes,

$$\left[\frac{\mu_{Y|j+1,k,G} - \mu_{Y|j-1,k,G}}{T_G(Matching1_{j+1}) - T_G(Matching1_{j-1})}\right] \text{ and } \left[\frac{\mu_{Y|j,k+1,G} - \mu_{Y|j,k-1,G}}{T_G(Matching2_{k+1}) - T_G(Matching2_{k-1})}\right],$$

which are allowed to vary at each level of *j* (conditional on *k*) and *k* (conditional on *j*). As with the 10 previously described DIF methods and matching variables, bivariate DIF estimates based on Equations 4 and 5 were computed after the frequency data were smoothed using loglinear models.

## Tests and Testing Programs Considered

The CR items from two forms of three mixed-format tests were evaluated for DIF with respect to gender, where females made up the focal groups and males made up the reference groups.

**SAT Math tests**. Two recent administrations of the SAT Math test were assessed. From each administration's test, 10 dichotomously scored student-produced response (SPR) items were assessed for CR DIF. The tests were also composed of 44 MC questions. The descriptive statistics for the administrations' test forms, anonymously labeled as Forms 1 and 2 in this report, are summarized in Tables 2 and 3.

**Table 2**

***Descriptive Statistics of the Data From the SAT Math Test, Form 1***

| Group | Score | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Males | SPR1 | 0 | 1 | 0.80 | 0.40 |
| (*N* = 204,956) | SPR2 | 0 | 1 | 0.56 | 0.50 |
| | SPR3 | 0 | 1 | 0.72 | 0.45 |
| | SPR4 | 0 | 1 | 0.58 | 0.49 |
| | SPR5 | 0 | 1 | 0.48 | 0.50 |
| | SPR6 | 0 | 1 | 0.29 | 0.45 |
| | SPR7 | 0 | 1 | 0.34 | 0.47 |
| | SPR8 | 0 | 1 | 0.27 | 0.44 |
| | SPR9 | 0 | 1 | 0.20 | 0.40 |
| | SPR10 | 0 | 1 | 0.14 | 0.35 |
| | MC | -7 | 44 | 24.38 | 10.38 |
| | CR | 0 | 10 | 4.37 | 2.61 |
| | MC + CR | -7 | 54 | 28.75 | 12.62 |
| Females | SPR1 | 0 | 1 | 0.75 | 0.43 |
| (*N* = 235,756) | SPR2 | 0 | 1 | 0.49 | 0.50 |
| | SPR3 | 0 | 1 | 0.62 | 0.49 |
| | SPR4 | 0 | 1 | 0.54 | 0.50 |
| | SPR5 | 0 | 1 | 0.42 | 0.49 |
| | SPR6 | 0 | 1 | 0.24 | 0.43 |
| | SPR7 | 0 | 1 | 0.27 | 0.44 |
| | SPR8 | 0 | 1 | 0.22 | 0.41 |
| | SPR9 | 0 | 1 | 0.12 | 0.33 |
| | SPR10 | 0 | 1 | 0.09 | 0.28 |
| | MC | -8 | 44 | 21.50 | 9.73 |
| | CR | 0 | 10 | 3.76 | 2.39 |
| | MC + CR | -8 | 54 | 25.26 | 11.74 |

*Note.* CR = constructed response; MC = multiple choice; SPR = student-produced response.

**Table 3**

***Descriptive Statistics of the Data From the SAT Math Test, Form 2***

| Group | Score | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Males | SPR1 | 0 | 1 | 0.74 | 0.44 |
| (*N* = 229,251) | SPR2 | 0 | 1 | 0.60 | 0.49 |
| | SPR3 | 0 | 1 | 0.76 | 0.42 |
| | SPR4 | 0 | 1 | 0.83 | 0.37 |
| | SPR5 | 0 | 1 | 0.64 | 0.48 |
| | SPR6 | 0 | 1 | 0.44 | 0.50 |
| | SPR7 | 0 | 1 | 0.51 | 0.50 |
| | SPR8 | 0 | 1 | 0.33 | 0.47 |
| | SPR9 | 0 | 1 | 0.27 | 0.45 |
| | SPR10 | 0 | 1 | 0.22 | 0.42 |
| | MC | -8 | 44 | 26.47 | 10.05 |
| | CR | 0 | 10 | 5.35 | 2.67 |
| | MC + CR | -8 | 54 | 31.82 | 12.38 |
| Females | SPR1 | 0 | 1 | 0.59 | 0.49 |
| (*N* = 291,963) | SPR2 | 0 | 1 | 0.54 | 0.50 |
| | SPR3 | 0 | 1 | 0.68 | 0.46 |
| | SPR4 | 0 | 1 | 0.79 | 0.40 |
| | SPR5 | 0 | 1 | 0.58 | 0.49 |
| | SPR6 | 0 | 1 | 0.34 | 0.47 |
| | SPR7 | 0 | 1 | 0.37 | 0.48 |
| | SPR8 | 0 | 1 | 0.24 | 0.42 |
| | SPR9 | 0 | 1 | 0.19 | 0.39 |
| | SPR10 | 0 | 1 | 0.13 | 0.33 |
| | MC | -7 | 44 | 23.25 | 9.94 |
| | CR | 0 | 10 | 4.46 | 2.56 |
| | MC + CR | -7 | 54 | 27.71 | 12.14 |

*Note.* CR = constructed response; MC = multiple choice; SPR = student-produced response.

**Praxis tests.** Two recent forms of the Praxis Principles of Learning & Teaching: Grades 7–12 test were assessed. These forms, anonymously labeled as Forms 1 and 2, included twelve 4-point CR items (with possible ratings from 0 to 2 and a weight of 2) and 24 and 23 MC items (Tables 4–5). Two recent forms of the Praxis School Leaders Licensure Assessment were assessed. These forms, anonymously labeled as Forms 1 and 2, included seven 6-point CR items (with possible ratings from 0 to 3 scored by two raters) and 76 MC items (Tables 6–7).

**Table 4**

***Descriptive Statistics of the Data From the Praxis Principles of Learning & Teaching Test, Form 1***

| Group | Score | Min | Max | Mean | ***SD*** |
|---|---|---|---|---|---|
| Males | CR1 | 0 | 4 | 2.03 | 1.33 |
| (*N* = 1,588) | CR2 | 0 | 4 | 2.43 | 1.21 |
| | CR3 | 0 | 4 | 2.51 | 1.26 |
| | CR4 | 0 | 4 | 2.13 | 1.36 |
| | CR5 | 0 | 4 | 2.06 | 1.41 |
| | CR6 | 0 | 4 | 2.09 | 1.58 |
| | CR7 | 0 | 4 | 2.27 | 1.44 |
| | CR8 | 0 | 4 | 1.94 | 1.46 |
| | CR9 | 0 | 4 | 2.22 | 1.46 |
| | CR10 | 0 | 4 | 1.63 | 1.44 |
| | CR11 | 0 | 4 | 1.69 | 1.48 |
| | CR12 | 0 | 4 | 1.66 | 1.52 |
| | MC | 0 | 24 | 16.28 | 4.04 |
| | CR | 2 | 46 | 24.66 | 7.80 |
| | MC + CR | 4 | 67 | 40.94 | 10.17 |
| Females | CR1 | 0 | 4 | 2.21 | 1.28 |
| (*N* = 1,914) | CR2 | 0 | 4 | 2.72 | 1.22 |
| | CR3 | 0 | 4 | 2.76 | 1.23 |
| | CR4 | 0 | 4 | 2.42 | 1.38 |
| | CR5 | 0 | 4 | 2.53 | 1.37 |
| | CR6 | 0 | 4 | 2.18 | 1.58 |
| | CR7 | 0 | 4 | 2.47 | 1.41 |
| | CR8 | 0 | 4 | 2.26 | 1.43 |
| | CR9 | 0 | 4 | 2.45 | 1.42 |
| | CR10 | 0 | 4 | 2.08 | 1.45 |
| | CR11 | 0 | 4 | 1.92 | 1.50 |
| | CR12 | 0 | 4 | 2.04 | 1.56 |
| | MC | 0 | 24 | 17.30 | 3.84 |
| | CR | 0 | 48 | 28.04 | 7.71 |
| | MC + CR | 0 | 70 | 45.34 | 10.04 |

*Note.* CR = constructed response; MC = multiple choice.

**Table 5**

***Descriptive Statistics of the Data From the Praxis Principles of Learning & Teaching Test, Form 2***

| Group | Score | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Males ($N$ = 1,482) | CR1 | 0 | 4 | 2.55 | 1.25 |
| | CR2 | 0 | 4 | 2.46 | 1.32 |
| | CR3 | 0 | 4 | 2.25 | 1.36 |
| | CR4 | 0 | 4 | 1.96 | 1.33 |
| | CR5 | 0 | 4 | 1.86 | 1.32 |
| | CR6 | 0 | 4 | 2.36 | 1.38 |
| | CR7 | 0 | 4 | 1.94 | 1.41 |
| | CR8 | 0 | 4 | 2.47 | 1.31 |
| | CR9 | 0 | 4 | 2.14 | 1.42 |
| | CR10 | 0 | 4 | 2.12 | 1.31 |
| | CR11 | 0 | 4 | 1.57 | 1.44 |
| | CR12 | 0 | 4 | 1.71 | 1.48 |
| | MC | 0 | 23 | 15.12 | 3.57 |
| | CR | 2 | 46 | 25.41 | 7.82 |
| | MC + CR | 10 | 67 | 40.52 | 10.00 |
| Females ($N$ = 1,936) | CR1 | 0 | 4 | 2.86 | 1.22 |
| | CR2 | 0 | 4 | 2.72 | 1.30 |
| | CR3 | 0 | 4 | 2.58 | 1.32 |
| | CR4 | 0 | 4 | 2.34 | 1.32 |
| | CR5 | 0 | 4 | 2.19 | 1.34 |
| | CR6 | 0 | 4 | 2.53 | 1.39 |
| | CR7 | 0 | 4 | 2.10 | 1.44 |
| | CR8 | 0 | 4 | 2.77 | 1.29 |
| | CR9 | 0 | 4 | 2.45 | 1.41 |
| | CR10 | 0 | 4 | 2.40 | 1.26 |
| | CR11 | 0 | 4 | 1.92 | 1.44 |
| | CR12 | 0 | 4 | 2.05 | 1.49 |
| | MC | 0 | 23 | 16.19 | 3.19 |
| | CR | 0 | 48 | 28.93 | 7.85 |
| | MC + CR | 0 | 69 | 45.11 | 9.76 |

*Note.* CR = constructed response; MC = multiple choice.

**Table 6**

***Descriptive Statistics of the Data From the Praxis School Leaders Licensure Assessment, Form 1***

| Group | Score | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Males | CR1 | 0 | 6 | 3.99 | 1.40 |
| (*N* = 407) | CR2 | 0 | 6 | 3.26 | 1.78 |
| | CR3 | 0 | 6 | 3.94 | 1.60 |
| | CR4 | 0 | 6 | 3.91 | 1.64 |
| | CR5 | 0 | 6 | 3.37 | 1.70 |
| | CR6 | 0 | 6 | 4.01 | 1.93 |
| | CR7 | 0 | 6 | 3.32 | 1.91 |
| | MC | 37 | 73 | 57.34 | 5.86 |
| | CR | 4 | 33 | 21.03 | 5.50 |
| | MC + CR | 45 | 102 | 78.37 | 9.85 |
| Females | CR1 | 0 | 6 | 4.33 | 1.37 |
| (*N* = 776) | CR2 | 0 | 6 | 3.58 | 1.71 |
| | CR3 | 0 | 6 | 4.31 | 1.49 |
| | CR4 | 0 | 6 | 4.20 | 1.62 |
| | CR5 | 0 | 6 | 3.76 | 1.74 |
| | CR6 | 0 | 6 | 4.34 | 1.80 |
| | CR7 | 0 | 6 | 3.43 | 1.97 |
| | MC | 33 | 73 | 58.51 | 5.64 |
| | CR | 4 | 34 | 22.79 | 5.31 |
| | MC + CR | 38 | 103 | 81.29 | 9.31 |

*Note.* CR = constructed response; MC = multiple choice.

**Table 7**

***Descriptive Statistics of the Data From the Praxis School Leaders Licensure Assessment, Form 2***

| Group | Score | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Males | CR1 | 0 | 6 | 4.08 | 1.37 |
| (*N* = 1,048) | CR2 | 0 | 6 | 3.58 | 1.70 |
| | CR3 | 0 | 6 | 4.06 | 1.52 |
| | CR4 | 0 | 6 | 4.46 | 1.53 |
| | CR5 | 0 | 6 | 3.55 | 1.70 |
| | CR6 | 0 | 6 | 3.72 | 1.80 |
| | CR7 | 0 | 6 | 3.53 | 1.93 |
| | MC | 35 | 72 | 58.01 | 6.05 |
| | CR | 5 | 34 | 21.96 | 5.15 |
| | MC + CR | 44 | 101 | 79.97 | 9.67 |

| Group | Score | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| Females | CR1 | 0 | 6 | 4.28 | 1.39 |
| (*N* = 1,816) | CR2 | 0 | 6 | 3.78 | 1.68 |
| | CR3 | 0 | 6 | 4.27 | 1.46 |
| | CR4 | 0 | 6 | 4.54 | 1.57 |
| | CR5 | 0 | 6 | 3.84 | 1.67 |
| | CR6 | 0 | 6 | 3.86 | 1.76 |
| | CR7 | 0 | 6 | 3.81 | 1.85 |
| | MC | 32 | 72 | 58.60 | 6.07 |
| | CR | 3 | 34 | 23.08 | 4.94 |
| | MC + CR | 37 | 104 | 81.68 | 9.60 |

*Note.* CR = constructed response; MC = multiple choice.

## Results

The CR DIF results for the considered test forms are presented in Tables 8–13. These tables show the tests' characteristics expected to affect the DIF methods and results, including the reliabilities of the MC and CR - *Y* scores, the correlations of the MC and CR - *Y* scores with *Y*, and measures of *impact* (Dorans & Holland, 1993, pp. 36–38) computed as the differences in the focal and reference groups' means (*F*-*R*) divided by the standard deviation of the focal and reference groups' scores for *Y*, and also for the MC and CR - *Y* scores. The tables' mean DIF values show the average of the 14 methods' CR DIF values. The variabilities of the 14 methods' DIF values are shown as the deviation of each method's DIF value from the mean DIF value. In the tables, results are presented first for the methods using the observed and estimated true scores of the CR - *Y* and CR matching variables, then for the methods using the observed and estimated true scores of the summed MC + CR - *Y* and MC + CR scores as matching variables, then for the methods using the observed and estimated true scores of the bivariate (MC, CR - *Y*) and (MC, CR) matching variables, and finally for methods using the observed and estimated true scores of the MC scores as matching variables.

### SAT Math Test Results

The SAT Math test results are presented in Tables 8–9. The impact values on *Y*, MC, and CR - *Y* are all negative, indicating that on average males outperformed females on the major sections of the tests. The impact values on the MC and CR - *Y* scores are similar, suggesting that the MC and CR sections of the SAT Math tests measure similar constructs. The reliabilities of the test sections are summarized at the bottom of the tables, showing that the MC sections

reached a reliability of 0.90 whereas the reliability levels of the CR sections were 0.73 and 0.75. Most *Y* scores had a higher correlation with the MC scores than with the CR scores.

Tables 8–9 show that the 14 DIF methods all produced small deviations from the mean DIF values. Some slight patterns in the results can be observed, in that negative deviations usually resulted from DIF methods that used MC as the matching variable, whereas positive deviations resulted from using *T*(CR) as the matching variable. These deviation patterns were relatively small and less distinct than those observed in the DIF results for the Praxis tests.

**Praxis Test Results**

The test characteristics and CR DIF results for the Praxis Principles of Learning & Teaching test and for the Praxis School Leaders Licensure Assessment are presented in Tables 10–13. The Praxis test results differ from those of the SAT Math test results with respect to overall test characteristics and the overall pattern of CR DIF results. In terms of test characteristics, Tables 10–13 indicate that while females generally outperformed males on both sections of the tests, these performance differences were greater on the CR - *Y* sections' scores than on the MC sections' scores. The impact values of the studied items were also positive and were usually more similar to those of the MC matching variable than the CR - *Y* matching variables. Compared to the SAT Math tests, the Praxis tests' MC and CR sections were less reliable and exhibited lower correlations for the studied items with the MC and CR - *Y* matching variables.

Tables 10–13 show that the Praxis tests have a consistent pattern of DIF results that differs from those of the SAT Math tests. The *T*(CR), *T*(CR - *Y*), CR, *T*(MC + CR), and [*T*(MC), *T*(CR)] matching variables produced DIF values with negative deviations from the mean DIF values. The MC matching variable produced DIF values with the largest positive deviations from the mean DIF values. Other matching variables that resulted in DIF values with positive deviations included *T*(MC), (MC, CR - *Y*), [*T*(MC), *T*(CR - *Y*)], MC + CR – *Y*, and (MC,CR). DIF results from the CR - *Y*, *T*(MC + CR - *Y*), and MC + CR matching variables had relatively small deviations from the mean DIF values.

**Table 8**

***Constructed-Response DIF Results for the SAT Math Test, Form 1***

| *F-R* impact on the studied item (*Y*) | *F-R* impact on the matching variables (MC, CR - *Y*) | Corr with *Y* (MC, CR - *Y*) | Mean DIF value | CR DIF results based on the following matching variables (deviations from the mean DIF value) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *T*(CR - *Y*) | CR - *Y* | *T*(CR) | CR | *T*(MC + CR - *Y*) | MC + CR - *Y* | *T*(MC + CR) | MC + CR | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* | *T*(MC), *T*(CR) | MC, CR | *T*(MC) | MC |
| SPR1, -0.12 | -0.29, -0.25 | 0.38, 0.35 | 0.00 | 0.00 | -0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.02 | -0.00 | 0.00 | -0.01 | 0.00 | 0.00 |
| SPR2, -0.14 | -0.29, -0.25 | 0.46, 0.41 | 0.01 | -0.00 | -0.02 | 0.01 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.01 | 0.00 | -0.01 |
| SPR3, -0.21 | -0.29, -0.23 | 0.55, 0.48 | 0.04 | 0.00 | -0.02 | 0.01 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 |
| SPR4, -0.08 | -0.29, -0.26 | 0.55, 0.49 | 0.05 | 0.00 | -0.02 | 0.01 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.00 | 0.02 | 0.01 | 0.00 | -0.01 |
| SPR5, -0.12 | -0.29, -0.24 | 0.50, 0.47 | 0.02 | 0.00 | -0.02 | 0.01 | -0.01 | 0.00 | -0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.03 | 0.01 | 0.00 | -0.01 |
| SPR6, -0.11 | -0.29, -0.24 | 0.28, 0.27 | 0.01 | -0.00 | -0.02 | 0.01 | 0.00 | -0.00 | -0.01 | 0.00 | -0.00 | 0.00 | -0.00 | 0.02 | 0.00 | -0.00 | -0.01 |
| SPR7, -0.15 | -0.29, -0.24 | 0.40, 0.38 | 0.01 | -0.00 | -0.03 | 0.01 | -0.01 | -0.00 | -0.01 | 0.00 | -0.00 | 0.01 | 0.00 | 0.03 | 0.01 | -0.00 | -0.01 |
| SPR8, -0.12 | -0.29, -0.25 | 0.54, 0.49 | 0.04 | 0.00 | -0.02 | 0.01 | -0.01 | 0.00 | -0.01 | 0.00 | -0.00 | 0.01 | 0.00 | 0.03 | 0.00 | 0.00 | -0.01 |
| SPR9, -0.22 | -0.29, -0.23 | 0.36, 0.34 | 0.02 | 0.00 | -0.02 | 0.01 | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.02 | 0.00 | 0.04 | 0.01 | 0.00 | -0.01 |
| SPR10, -0.16 | -0.29, -0.23 | 0.35, 0.34 | 0.00 | -0.00 | -0.02 | 0.01 | -0.01 | -0.00 | -0.01 | -0.00 | -0.00 | 0.00 | 0.00 | 0.03 | 0.00 | -0.00 | -0.01 |

*Note.* The reliabilities of matching variables MC and CR were approximately 0.90 and 0.73, respectively. Corr = correlated; CR = constructed response; DIF = differential item functioning; F-R = focal-reference; MC = multiple choice; SPR = student-produced response.

**Table 9**

***Constructed-Response DIF Results for the SAT Math Test, Form 2***

| *F-R* impact on the studied item (*Y*) | *F-R* impact on the matching variables (MC, CR - *Y*) | Corr with *Y* (MC, CR - *Y*) | Mean DIF value | CR DIF results based on the following matching variables (deviations from the mean DIF value) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *T*(CR - *Y*) | CR - *Y* | *T*(CR) | CR | *T*(MC + CR - *Y*) | MC + CR - *Y* | *T*(MC + CR) | MC + CR | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* | *T*(MC), *T*(CR) | MC, CR | *T*(MC) | MC |
| SPR1, -0.32 | -0.32, -0.31 | 0.44, 0.41 | -0.07 | 0.00 | -0.02 | 0.03 | 0.00 | -0.01 | -0.01 | 0.00 | -0.01 | 0.00 | 0.00 | 0.01 | 0.00 | -0.01 | -0.02 |
| SPR2, -0.12 | -0.32, -0.36 | 0.54, 0.47 | 0.04 | 0.01 | -0.01 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.01 | 0.03 | 0.01 | 0.00 | -0.01 |
| SPR3, -0.18 | -0.32, -0.35 | 0.54, 0.47 | 0.00 | 0.01 | -0.01 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.01 | 0.02 | 0.00 | 0.00 | 0.00 |
| SPR4, -0.10 | -0.32, -0.35 | 0.31, 0.29 | 0.00 | 0.01 | -0.01 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | -0.02 | -0.02 | 0.00 | -0.01 | 0.00 | -0.01 |
| SPR5, -0.12 | -0.32, -0.36 | 0.55, 0.48 | 0.03 | 0.01 | -0.01 | 0.03 | 0.01 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | -0.01 | 0.02 | 0.00 | 0.00 | -0.01 |
| SPR6, -0.21 | -0.32, -0.34 | 0.46, 0.44 | 0.00 | 0.01 | -0.02 | 0.03 | 0.01 | -0.01 | -0.01 | 0.00 | -0.01 | 0.01 | 0.00 | 0.03 | 0.01 | -0.01 | -0.02 |
| SPR7, -0.29 | -0.32, -0.32 | 0.55, 0.52 | -0.03 | 0.00 | -0.02 | 0.03 | 0.00 | -0.01 | -0.01 | 0.00 | -0.01 | 0.02 | 0.00 | 0.03 | 0.01 | -0.01 | -0.02 |
| SPR8, -0.20 | -0.32, -0.33 | 0.36, 0.36 | -0.01 | 0.00 | -0.02 | 0.02 | 0.00 | -0.01 | -0.02 | -0.00 | -0.01 | 0.01 | 0.00 | 0.03 | 0.01 | -0.01 | -0.02 |
| SPR9, -0.19 | -0.32, -0.35 | 0.50, 0.48 | 0.02 | 0.00 | -0.02 | 0.02 | -0.01 | -0.01 | -0.01 | 0.00 | -0.01 | 0.01 | 0.00 | 0.03 | 0.00 | -0.01 | -0.02 |
| SPR10, -0.24 | -0.32, -0.33 | 0.44, 0.44 | -0.01 | 0.00 | -0.02 | 0.01 | 0.00 | -0.01 | -0.01 | 0.00 | -0.01 | 0.01 | 0.00 | 0.03 | 0.00 | -0.01 | -0.02 |

*Note.* The reliabilities of matching variables MC and CR were approximately 0.90 and 0.75, respectively. F-R = focal-reference; Corr = correlated; CR = constructed response; DIF = differential item functioning; MC = multiple choice; SPR = student-produced response.

**Table 10**

***Constructed-Response DIF Results for the Praxis Principles of Learning & Teaching Test, Form 1***

| *F-R* impact on the studied item (*Y*) | *F-R* impact on the matching variables (MC, CR - *Y*) | Corr with *Y* (MC, CR - *Y*) | Mean DIF value | CR DIF results based on the following matching variables (deviations from the mean DIF value) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *T*(CR - *Y*) | CR - *Y* | *T*(CR) | CR | *T*(MC + CR - *Y*) | MC + CR - *Y* | *T*(MC + CR) | MC + CR | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* | *T*(MC), *T*(CR) | MC, CR | *T*(MC) | MC |
| CR1, 0.14 | 0.26, 0.42 | 0.16, 0.17 | 0.05 | 0.00 | 0.05 | -0.15 | -0.05 | -0.01 | 0.03 | -0.11 | -0.04 | 0.06 | 0.08 | -0.06 | 0.05 | 0.06 | 0.08 |
| CR2, 0.24 | 0.26, 0.41 | 0.17, 0.22 | 0.14 | -0.02 | 0.04 | -0.14 | -0.04 | -0.01 | 0.03 | -0.10 | -0.04 | 0.05 | 0.07 | -0.05 | 0.05 | 0.06 | 0.08 |
| CR3, 0.20 | 0.26, 0.42 | 0.20, 0.27 | 0.11 | -0.04 | 0.03 | -0.17 | -0.06 | -0.02 | 0.03 | -0.10 | -0.04 | 0.06 | 0.09 | -0.03 | 0.06 | 0.08 | 0.10 |
| CR4, 0.21 | 0.26, 0.42 | 0.18, 0.25 | 0.10 | -0.04 | 0.04 | -0.19 | -0.06 | -0.02 | 0.04 | -0.12 | -0.04 | 0.06 | 0.09 | -0.05 | 0.06 | 0.10 | 0.12 |
| CR5, 0.34 | 0.26, 0.39 | 0.21, 0.25 | 0.27 | -0.03 | 0.05 | -0.20 | -0.06 | -0.02 | 0.04 | -0.13 | -0.05 | 0.05 | 0.09 | -0.06 | 0.05 | 0.09 | 0.13 |
| CR6, 0.06 | 0.26, 0.44 | 0.11, 0.20 | -0.10 | -0.04 | 0.03 | -0.24 | -0.10 | -0.01 | 0.03 | -0.15 | -0.06 | 0.09 | 0.12 | -0.04 | 0.07 | 0.12 | 0.14 |
| CR7, 0.14 | 0.26, 0.43 | 0.15, 0.27 | 0.03 | -0.06 | 0.03 | -0.21 | -0.08 | -0.02 | 0.03 | -0.13 | -0.05 | 0.08 | 0.11 | -0.04 | 0.07 | 0.11 | 0.14 |
| CR8, 0.22 | 0.26, 0.42 | 0.23, 0.31 | 0.06 | -0.05 | 0.06 | -0.21 | -0.06 | -0.04 | 0.04 | -0.15 | -0.05 | 0.07 | 0.12 | -0.05 | 0.08 | 0.13 | 0.16 |
| CR9, 0.16 | 0.26, 0.44 | 0.28, 0.41 | -0.06 | -0.12 | 0.04 | -0.23 | -0.05 | -0.08 | 0.02 | -0.16 | -0.05 | 0.07 | 0.13 | -0.03 | 0.10 | 0.14 | 0.18 |
| CR10, 0.31 | 0.26, 0.41 | 0.28, 0.41 | 0.16 | -0.09 | 0.06 | -0.24 | -0.06 | -0.05 | 0.04 | -0.17 | -0.05 | 0.07 | 0.13 | -0.05 | 0.10 | 0.16 | 0.20 |
| CR11, 0.15 | 0.26, 0.44 | 0.24, 0.38 | -0.07 | -0.13 | 0.03 | -0.26 | -0.07 | -0.06 | 0.04 | -0.17 | -0.05 | 0.09 | 0.15 | -0.02 | 0.12 | 0.18 | 0.21 |
| CR12, 0.25 | 0.26, 0.42 | 0.25, 0.38 | 0.07 | -0.10 | 0.05 | -0.26 | -0.07 | -0.06 | 0.04 | -0.18 | -0.06 | 0.07 | 0.13 | -0.05 | 0.09 | 0.16 | 0.20 |

*Note.* The reliabilities of matching variables MC and CR were approximately 0.74 and 0.64, respectively. Corr = correlated; CR = constructed response; DIF = differential item functioning; F-R = focal-reference; MC = multiple choice.

**Table 11**

***Constructed-Response DIF Results for the Praxis Principles of Learning & Teaching Test, Form 2***

| *F-R* impact on the studied item (*Y*) | *F-R* impact on the matching variables (MC, CR - *Y*) | Corr with *Y* (MC, CR - *Y*) | Mean DIF value | CR DIF results based on the following matching variables (deviations from the mean DIF value) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *T*(CR - *Y*) | CR - *Y* | *T*(CR) | CR | *T*(MC + CR - *Y*) | MC + CR - *Y* | *T*(MC + CR) | MC + CR | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* | *T*(MC), *T*(CR) | MC, CR | *T*(MC) | MC |
| CR1, 0.25 | 0.31, 0.42 | 0.19, 0.24 | 0.15 | 0.00 | 0.05 | -0.13 | -0.04 | -0.01 | 0.03 | -0.11 | -0.04 | 0.03 | 0.07 | -0.06 | 0.05 | 0.05 | 0.08 |
| CR2, 0.20 | 0.31, 0.43 | 0.22, 0.33 | 0.07 | -0.02 | 0.05 | -0.15 | -0.04 | -0.02 | 0.04 | -0.12 | -0.04 | 0.04 | 0.09 | -0.06 | 0.06 | 0.07 | 0.11 |
| CR3, 0.25 | 0.31, 0.42 | 0.16, 0.27 | 0.13 | -0.05 | 0.03 | -0.19 | -0.07 | -0.02 | 0.04 | -0.13 | -0.05 | 0.07 | 0.11 | -0.04 | 0.08 | 0.11 | 0.14 |
| CR4, 0.29 | 0.31, 0.42 | 0.32, 0.38 | 0.10 | -0.05 | 0.06 | -0.18 | -0.04 | -0.06 | 0.03 | -0.15 | -0.04 | 0.03 | 0.11 | -0.08 | 0.08 | 0.09 | 0.15 |
| CR5, 0.25 | 0.31, 0.43 | 0.17, 0.30 | 0.11 | -0.04 | 0.05 | -0.19 | -0.06 | -0.02 | 0.04 | -0.14 | -0.05 | 0.06 | 0.11 | -0.06 | 0.08 | 0.11 | 0.14 |
| CR6, 0.12 | 0.31, 0.45 | 0.19, 0.31 | -0.08 | -0.06 | 0.04 | -0.18 | -0.05 | -0.05 | 0.03 | -0.14 | -0.04 | 0.07 | 0.12 | -0.03 | 0.09 | 0.11 | 0.15 |
| CR7, 0.11 | 0.31, 0.45 | 0.22, 0.29 | -0.04 | -0.02 | 0.05 | -0.18 | -0.06 | -0.02 | 0.04 | -0.14 | -0.05 | 0.04 | 0.10 | -0.08 | 0.06 | 0.08 | 0.12 |
| CR8, 0.23 | 0.31, 0.44 | 0.27, 0.39 | 0.07 | -0.06 | 0.03 | -0.17 | -0.05 | -0.05 | 0.02 | -0.14 | -0.04 | 0.05 | 0.11 | -0.04 | 0.08 | 0.09 | 0.14 |
| CR9, 0.22 | 0.31, 0.44 | 0.25, 0.39 | 0.02 | -0.08 | 0.04 | -0.20 | -0.06 | -0.07 | 0.02 | -0.16 | -0.05 | 0.06 | 0.13 | -0.05 | 0.10 | 0.12 | 0.17 |
| CR10, 0.22 | 0.31, 0.44 | 0.26, 0.41 | 0.01 | -0.06 | 0.05 | -0.17 | -0.04 | -0.06 | 0.03 | -0.14 | -0.04 | 0.05 | 0.12 | -0.05 | 0.09 | 0.11 | 0.16 |
| CR11, 0.24 | 0.31, 0.43 | 0.26, 0.37 | 0.05 | -0.05 | 0.06 | -0.22 | -0.06 | -0.05 | 0.04 | -0.17 | -0.05 | 0.05 | 0.13 | -0.07 | 0.09 | 0.12 | 0.18 |
| CR12, 0.23 | 0.31, 0.43 | 0.22, 0.35 | 0.03 | -0.04 | 0.07 | -0.21 | -0.05 | -0.05 | 0.04 | -0.17 | -0.05 | 0.06 | 0.13 | -0.07 | 0.09 | 0.13 | 0.18 |

*Note.* The reliabilities of matching variables MC and CR were approximately 0.64 and 0.68, respectively. Corr = correlated; CR = constructed response; DIF = differential item functioning; F-R = focal-reference; MC = multiple choice.

**Table 12**

***Constructed-Response DIF Results for the Praxis School Leaders Licensure Assessment, Form 1***

| *F-R* impact on the studied item (*Y*) | *F-R* impact on the matching variables (MC, CR - *Y*) | Corr with *Y* (MC, CR - *Y*) | Mean DIF value | CR DIF results based on the following matching variables (deviations from the mean DIF value) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *T*(CR - *Y*) | CR - *Y* | *T*(CR) | CR | *T*(MC + CR - *Y*) | MC + CR - *Y* | *T*(MC + CR) | MC + CR | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* | *T*(MC), *T*(CR) | MC, CR | *T*(MC) | MC |
| CR1, 0.25 | 0.20, 0.29 | 0.13, 0.13 | 0.22 | 0.02 | 0.06 | -0.13 | -0.03 | 0.01 | 0.04 | -0.05 | -0.01 | 0.02 | 0.04 | -0.06 | 0.03 | 0.04 | 0.06 |
| CR2, 0.18 | 0.20, 0.30 | 0.27, 0.25 | 0.06 | -0.04 | 0.08 | -0.26 | -0.06 | -0.01 | 0.06 | -0.11 | -0.02 | 0.04 | 0.10 | -0.06 | 0.08 | 0.09 | 0.14 |
| CR3, 0.24 | 0.20, 0.29 | 0.22, 0.29 | 0.17 | -0.05 | 0.05 | -0.22 | -0.06 | 0.00 | 0.05 | -0.08 | -0.01 | 0.03 | 0.08 | -0.07 | 0.05 | 0.08 | 0.11 |
| CR4, 0.18 | 0.20, 0.31 | 0.27, 0.34 | 0.08 | -0.09 | 0.04 | -0.26 | -0.07 | 0.00 | 0.06 | -0.08 | -0.01 | 0.05 | 0.10 | -0.05 | 0.08 | 0.10 | 0.14 |
| CR5, 0.23 | 0.20, 0.30 | 0.29, 0.41 | 0.17 | -0.08 | 0.07 | -0.26 | -0.06 | 0.00 | 0.07 | -0.10 | -0.01 | 0.04 | 0.10 | -0.06 | 0.08 | 0.09 | 0.13 |
| CR6, 0.18 | 0.20, 0.31 | 0.22, 0.39 | 0.06 | -0.15 | 0.04 | -0.34 | -0.11 | 0.01 | 0.07 | -0.12 | -0.03 | 0.09 | 0.15 | -0.04 | 0.11 | 0.14 | 0.18 |
| CR7, 0.06 | 0.20, 0.36 | 0.28, 0.35 | -0.21 | -0.16 | 0.06 | -0.35 | -0.09 | -0.03 | 0.07 | -0.13 | -0.02 | 0.08 | 0.16 | -0.04 | 0.13 | 0.13 | 0.19 |

*Note.* The reliabilities of matching variables MC and CR were approximately 0.67 and 0.57, respectively. Corr = correlated; CR = constructed response; DIF = differential item functioning; F-R = focal-reference; MC = multiple choice.

**Table 13**

***Constructed-Response DIF Results for the Praxis School Leaders Licensure Assessment, Form 2***

| *F-R* impact on the studied item (*Y*) | *F-R* impact on the matching variables (MC, CR - *Y*) | Corr with *Y* (MC, CR - *Y*) | Mean DIF value | CR DIF results based on the following matching variables (deviations from the mean DIF value) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *T*(CR - *Y*) | CR - *Y* | *T*(CR) | CR | *T*(MC + CR - *Y*) | MC + CR - *Y* | *T*(MC + CR) | MC + CR | *T*(MC), *T*(CR - *Y*) | MC, CR - *Y* | *T*(MC), *T*(CR) | MC, CR | *T*(MC) | MC |
| CR1, 0.14 | 0.10, 0.21 | 0.20, 0.17 | 0.13 | 0.01 | 0.04 | -0.11 | -0.02 | 0.01 | 0.03 | -0.03 | 0.00 | 0.03 | 0.04 | -0.04 | 0.03 | 0.03 | 0.04 |
| CR2, 0.12 | 0.10, 0.22 | 0.33, 0.26 | 0.05 | -0.06 | 0.04 | -0.20 | -0.05 | -0.01 | 0.03 | -0.06 | 0.00 | 0.04 | 0.07 | -0.03 | 0.06 | 0.07 | 0.09 |
| CR3, 0.14 | 0.10, 0.21 | 0.24, 0.23 | 0.10 | -0.03 | 0.03 | -0.16 | -0.05 | 0.01 | 0.03 | -0.04 | 0.00 | 0.04 | 0.05 | -0.03 | 0.04 | 0.06 | 0.07 |
| CR4, 0.05 | 0.10, 0.24 | 0.27, 0.27 | -0.02 | -0.07 | 0.02 | -0.17 | -0.05 | 0.01 | 0.03 | -0.03 | 0.00 | 0.04 | 0.06 | -0.04 | 0.04 | 0.05 | 0.07 |
| CR5, 0.17 | 0.10, 0.20 | 0.15, 0.27 | 0.18 | -0.06 | 0.02 | -0.22 | -0.08 | 0.03 | 0.05 | -0.03 | 0.00 | 0.05 | 0.06 | -0.03 | 0.04 | 0.08 | 0.09 |
| CR6, 0.08 | 0.10, 0.24 | 0.32, 0.35 | -0.05 | -0.17 | 0.01 | -0.27 | -0.07 | 0.01 | 0.05 | -0.05 | 0.01 | 0.08 | 0.11 | -0.01 | 0.09 | 0.11 | 0.13 |
| CR7, 0.15 | 0.10, 0.21 | 0.35, 0.36 | 0.10 | -0.14 | 0.03 | -0.31 | -0.10 | 0.03 | 0.07 | -0.06 | 0.00 | 0.07 | 0.10 | -0.02 | 0.08 | 0.11 | 0.14 |

*Note.* The reliabilities of matching variables MC and CR were approximately 0.71 and 0.51, respectively. Corr = correlated; CR = constructed response; DIF = differential item functioning; F-R = focal-reference; MC = multiple choice.

## Discussion

CR DIF evaluations for mixed-format tests can be difficult to implement due to ambiguities about which DIF methods and matching variables are most appropriate. Surveys of ETS statistical coordinators suggest that CR DIF ambiguities may be reasons why CR DIF evaluations are not routinely conducted in the majority of ETS testing programs. The analyses and results in this paper demonstrate the complexities of CR DIF, suggesting that different DIF results could be obtained for different types of mixed-format tests and from using different matching variables and DIF methods.

Distinct patterns were visible in this study's CR DIF results based on the characteristics of the mixed-format tests. The pattern of CR DIF results for the student-produced CR items of the SAT Math test showed little to no variations among the 14 considered DIF methods' mean deviations. The Praxis tests' DIF results showed more variation among the methods, where the most negative DIF results were obtained using the *T*(CR), *T*(CR - *Y*), CR, *T*[MC + CR], and [*T*(MC), *T*(CR)] matching variables, and the most positive DIF results were obtained using the MC, *T*(MC), (MC, CR - *Y*), [*T*(MC), *T*(CR - *Y*)], MC + CR - *Y*, and (MC, CR) matching variables.

Although the SAT and Praxis tests differed with respect to several characteristics, the characteristics most aligned with these tests' CR DIF results appeared to be measures of impact on the potential matching variables, MC and CR - *Y*, and on *Y*. For the SAT Math test, the impact measures on the MC and CR - *Y* scores were relatively similar, suggesting that either score would produce similar results when used as a DIF matching variable. For the Praxis tests, the impact measures were positive on the MC scores and were more extremely positive on the CR - *Y* scores, resulting in a more complex pattern of DIF results. CR DIF evaluations for other mixed-format tests not described in this study produced additional patterns of impact and CR DIF results, where negative (positive) impact values on CR - *Y* (MC) matching variables resulted in negative (positive) DIF deviations when the CR - *Y* (MC) scores were used as matching variables. The suggestion that measures of impact might be useful in accounting for patterns of CR DIF results based on different DIF matching variables is a possible basis for future research and practice.

## Implications for CR DIF Research

Prior research about CR DIF has often considered issues such as the reliability of the matching variable, the use of total test scores as matching variables, and the implications of including the studied item in the total test score matching variable (Chang et al., 1996; Dorans & Schmitt, 1993; Kim et al., 2007; Kristjansson et al., 2005; Penfield, 2007; Penfield & Algina, 2006; Zwick et al., 1993; Zwick et al., 1997). The current study's results suggest that research should also consider measures of impact in the MC and CR section scores of mixed-format tests. Impact measures for section scores might indicate the MC and CR scores' measurement similarity to each other and to the studied item and might also indicate the scores' usefulness as potential DIF matching variables. Future research might utilize analyses and presentations like those in this study to evaluate the usefulness of impact measures with respect to characteristics like reliabilities and studied item correlations as bases for determining the most appropriate DIF estimate. These potential research studies could consider the best ways to use measures of impact and test characteristics to interpret the estimates of several CR DIF methods and matching variables obtained from a range of simulated and systematically manipulated conditions. Simulation studies would support evaluations of DIF methods' accuracies, evaluations of which were not possible with the current study's empirical analyses. Simulations could also inform the development of flagging rules for identifying situations where particular CR DIF methods and matching variables may be problematic and not advisable.

## Implications and Recommendations for Constructed-Response DIF Practice

The complexities of CR DIF evaluations are likely to be high for most mixed-format test data encountered in practice, where tests' MC and CR scores can vary in their measurement homogeneity, reliabilities, and the extent to which these scores reflect subgroup impact. One recommendation for addressing these complexities is to use this study's analyses and results tables to consider CR DIF results with respect to multiple matching variables and also with respect to test characteristics. This study's analysis presentations are useful for identifying situations where MC and CR scores are relatively similar and produce similar DIF results (e.g., the SAT tests) and other situations where MC and CR scores differ enough to warrant a choice of the most appropriate matching variable (e.g., the Praxis tests). As in current practice, the choices for addressing heterogeneous CR DIF results require judgments about the matching variables. This study's analysis presentations can inform judgments about matching variables and DIF

results because the presentations facilitate the assessment and interpretation of test characteristics on DIF results, including the effects of relatively unreliable CR section scores, of more reliable but less similar MC scores, of the use of summed or bivariate MC and CR scores to produce less extreme DIF results than the use of either MC or CR scores, and of inclusion or exclusion of the studied item. From prior research, current CR DIF practice at ETS, and the results of this study, the most recommendable matching variables are those with measurement characteristics that resemble the total test and the studied item. Based on the current study's results and analysis presentations, impact measures and comparative DIF presentations can be used to evaluate the measurement similarity of the studied item and the potential matching variables and to gauge the appropriateness and implications of potential DIF matching variables and CR DIF methods' results.

## References


Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–353.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

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

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133–183.

Kelley, T. L. (1923). *Statistical methods.* New York, NY: Macmillan.

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*, 93–116.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*, 935–953.

Moses, T., Miao, J., & Dorans, N. J. (2010). A comparison of strategies for estimating conditional DIF. *Journal of Educational and Behavioral Statistics*, *35*, 726–743.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* Fort Worth, TX: Harcourt Brace.

Penfield, R. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*, 187–210.

Penfield, R., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*, 295–312.

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

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items* (Research Report No. RR-97-05). Princeton, NJ: Educational Testing Service.