Compendium Study

Constructed-Response (CR) Differential Item Functioning (DIF) Evaluations for the *TOEIC*® Speaking and Writing Tests

Tim Moses

September 2013

ETS TOEIC

Know English. Know Success.

Recent and informal surveys of ETS statistical staff working with the data of various testing programs have indicated that, like the *TOEIC*® testing program, most testing programs administering constructed-response (CR) items do not routinely evaluate those items for differential item functioning (DIF). One of the reasons given for the avoidance of CR DIF evaluations across ETS testing programs is a lack of clarity about which matching variable and DIF method to use. An attempt to address the uncertainties of ETS statistical staff was made by Moses, Liu, Tan, Deng, and Dorans (in press). Moses et al.'s exploratory study conducted CR DIF evaluations of non-TOEIC mixed format tests by applying and comparing 14 different methods recommended in the CR DIF literature (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). The considered methods included those utilizing matching variables based on observed or true score estimates of the tests' composite scores, CR scores and multiple-choice (MC) scores, and also matching variables that included or excluded the studied CR item being evaluated for DIF (*Y*).

Results from the Moses et al. (in press) study were test-specific, suggesting that CR DIF investigations utilizing several methods can produce more or less homogeneous CR DIF results depending on the psychometric characteristics of the test, the matching variables, and the studied items. That is, for mixed format tests with MC and CR scores with more psychometrically desirable measurement properties (i.e., MC and CR tests that measure similar constructs and are long and of high reliability), similar CR DIF results can be obtained using a range of DIF methods and matching variables. Conversely, for mixed format tests with less psychometrically desirable measurement properties (i.e., MC and/or CR tests that measure different constructs or are short and of low reliability), CR DIF results obtained from different methods and matching variables can vary more and require careful evaluations of the studied items and matching variables to select the most appropriate CR DIF method. Two implications of these results are that exploratory evaluations that focus on the similarity of the results of multiple CR DIF methods can be useful for informing the choice of DIF method and also for understanding how the psychometric characteristics of a test's scores and items affect the results of the CR DIF methods.

The current study extended the Moses et al. (in press) explorations by considering a subset of the study's methods applied to the Speaking and Writing tests of the TOEIC testing program (Educational Testing Service, 2010). Moses et al. noted that this subset of CR DIF methods are described and recommended in the CR DIF literature, are used by the ETS testing programs that actually conduct CR DIF evaluations, and are most appropriate for the CR-only *TOEIC*® Speaking and Writing tests. This subset contains CR DIF methods and matching variables that are based on the total CR scores of the tests and that either include or exclude the studied CR item being evaluated for DIF. The total CR scores are used in their observed form (i.e., the standardized E-DIF method, Dorans & Schmitt, 1993) and in an estimated true score form (i.e., the PolySIB version of the SIBTEST, Chang et al., 1996). This study was intended to be useful for informing and addressing the limited use of CR DIF evaluations in the TOEIC Speaking and Writing tests.

Method

The focus of this study was the application of four CR DIF methods considered in Moses et al. (in press) to a recent form of the TOEIC Speaking and Writing tests. After a description of the Speaking and Writing test items and scores, the four CR DIF methods are described, including two implementations of the standardized E-DIF method (Dorans & Schmitt, 1993) and two implementations of the PolySIB DIF method (Chang et al., 1996). The CR DIF implementations of the standardized E-DIF and PolySIB DIF methods are responsive to an initial survey of ETS testing programs that indicated that the non-TOEIC programs that routinely assess CR DIF use these DIF methods. The implementations of the standardized E-DIF and PolySIB methods are based on four matching variables, including observed score and estimated true score versions of the total scores of the CR tests and the CR scores of the tests after excluding the studied item (*CR - Y*).

TOEIC® Speaking and Writing Test Forms, Item Weights in the Weighted Total Test Score Calculations, and Examinee Data

The CR items from a recently administered form of the TOEIC Speaking and Writing tests were evaluated for DIF with respect to gender, where females made up the focal groups and males made up the reference groups. The *TOEIC*® Speaking test is composed of 11 CR items that provide 13 scores (described in Table 1 and in Educational Testing Service, 2010). The *TOEIC*® Writing test is composed of eight CR items, each of which produces a single score described in Table 2 and in Educational Testing Service (2010). Descriptive statistics for the items and weighted total scores for the male and female examinees taking the test form used in the analyses of the study are summarized in Tables 1 and 2.

Table 1

Descriptive Statistics for the TOEIC Speaking Test Form

Item	Male (*N* = 240)				Female (*N* = 436)			
	Min	Max	Mean	SD	Min	Max	Mean	SD
SCR1: Read a text aloud (pronunciation)	1	3	2.28	0.51	1	3	2.50	0.51
SCR1: Read a text aloud (intonation)	1	3	2.24	0.52	1	3	2.48	0.51
SCR2: Read a text aloud (pronunciation)	1	3	2.27	0.52	1	3	2.37	0.53
SCR2: Read a text aloud (intonation)	1	3	2.23	0.50	1	3	2.36	0.52
SCR3: Describe a picture	1	3	2.45	0.55	1	3	2.64	0.50
SCR4: Respond to questions	0	3	2.15	0.84	0	3	2.25	0.84
SCR5: Respond to questions	0	3	2.23	0.68	1	3	2.51	0.63
SCR6: Respond to questions	0	3	1.19	0.77	0	3	2.02	0.81
SCR7: Respond to questions using information provided	1	3	2.57	0.50	1	3	2.61	0.55
SCR8: Respond to questions using information provided	0	3	2.29	0.64	1	3	2.48	0.64
SCR9: Respond to questions using information provided	1	3	2.25	0.52	1	3	2.35	0.51
SCR10: Propose a solution	1	5	3.07	0.65	2	5	3.21	0.62
SCR11: Express an opinion	2	5	3.30	0.65	2	5	3.49	0.67
Total Speaking score (CR)	9	23	16.35	2.41	11	24	17.27	2.21

Note. CR = constructed response, SCR = TOEIC Speaking test constructed-response item.

Table 2

Descriptive Statistics for the TOEIC Writing Test Form

Item	Male (*N* = 240)				Female (*N* = 436)			
	Min	Max	Mean	SD	Min	Max	Mean	SD
WCR1: Write a sentence based on a picture	1	3	2.77	0.48	1	3	2.85	0.42
WCR2: Write a sentence based on a picture	1	3	2.60	0.53	0	3	2.73	0.49
WCR3: Write a sentence based on a picture	1	3	2.43	0.60	0	3	2.55	0.58
WCR4: Write a sentence based on a picture	0	3	2.23	0.68	0	3	2.43	0.65
WCR5: Write a sentence based on a picture	0	3	2.38	0.76	0	3	2.58	0.63
WCR6: Respond to a written request	1	4	3.30	0.69	1	4	3.56	0.62
WCR7: Respond to a written request	1	4	3.01	0.78	1	4	3.37	0.68
WCR8: Write an opinion essay	1	5	3.09	0.53	2	5	3.25	0.52
Total Writing score (CR)	9	26	18.06	2.52	12	26	19.37	2.32

Note. CR = constructed response, WCR = TOEIC Writing test constructed-response item.

The total CR scores of the TOEIC Speaking test are computed as a rounded and weighted sum of averages of three sets of the 13 scores from the 11 items on the test. The rationale for this computation is that each of the i = 1 to 13 CR scores are regarded as corresponding to one of *j* = 1 to 3 claims about the speaking ability of test takers as nonnative English speakers to use spoken English in the context of everyday life and the global workplace. The five scores of the first three items correspond to Claim 1, which addresses relatively easy tasks that provide information about a test taker's ability to produce language that is intelligible to native and proficient nonnative English speakers. The next five scores of Items 4–9 correspond to Claim 2, which addresses tasks of intermediate difficulty that provide information about a test-taker's ability to carry out routine social and occupational interactions such as giving and receiving directions, asking for information, or asking for clarification. The final two scores of Items 10 and 11 correspond to Claim 3, which addresses the most difficult tasks that provide information about a test-taker's ability to create connected and sustained discourse appropriate to the typical workplace.

Using the notation and descriptions provided above, the computation of the total CR scores of the TOEIC Speaking test can be expressed as

$$\begin{aligned} CR_{\text{Speaking}} &= \text{round}[\sum_{j}(\text{Claim}_j)\frac{1}{\#\text{scores}_j}(\sum_{i\ \text{in}\ j}\text{SCR}_i)] \\ &= \text{round}[(1)\frac{1}{5}(\text{SCR1}_{\text{Pronounciation}}+\text{SCR1}_{\text{Intonation}}+\text{SCR2}_{\text{Pronounciation}}+\text{SCR2}_{\text{Intonation}}+\text{SCR3}) \\ &\quad +(2)\frac{1}{6}(\text{SCR4}+\text{SCR5}+\text{SCR6}+\text{SCR7}+\text{SCR8}+\text{SCR9}) \\ &\quad +(3)\frac{1}{2}(\text{SCR10}+\text{SCR11})] \end{aligned}, \quad (1)$$

where the *j*th claim's number, Claim_j, is also the weight given to that claim's average scores, $\frac{1}{\#\text{scores}_j}(\sum_{i\ \text{in}\ j}\text{SCR}_i)$, in the computation of the total CR score of the TOEIC Speaking test.

The total CR scores of the TOEIC Writing test are computed as a rounded and weighted sum of the eight scores and items of the test. The rationale for this computation is that each of the *k* = 1 to 8 CR scores are regarded as corresponding to one of *j* = 1 to 3 claims about the ability of nonnative English speakers to use written English in the context of everyday life and the global workplace. The scores of first five items correspond to Claim 1, which addresses relatively easy tasks that provide information about a test taker's ability to produce well-formed sentences. The scores of the next two items, Items 6 and 7, correspond to Claim 2, which addresses tasks of intermediate difficulty that provide information about the ability of a test taker to produce multisentence length text to convey straightforward information, questions, instructions, narratives, and so forth. The final score of Item 8 corresponds to Claim 3, addressing the most difficult task that provides information about a test taker's ability to produce multiparagraph length text to express complex ideas, using, as appropriate, reasons, evidence, and extended explanations.

Using the notation and descriptions provided above, the computation of the total CR scores of the TOEIC Writing test can be expressed as

$$
\begin{aligned}
CR_{\text{Writing}} &= \text{round}[\sum_{j}(\text{Claim}_j)\frac{1}{\#\text{scores}_j}(\sum_{k\ \text{in}\ j}\text{WCR}_k)] \\
&= \text{round}[(1)\frac{1}{5}(\text{WCR1}+\text{WCR2}+\text{WCR3}+\text{WCR4}+\text{WCR5}) \\
&\quad +(2)\frac{1}{2}(\text{WCR6}+\text{WCR7}) \\
&\quad +(3)\frac{1}{1}*(\text{WCR8})]
\end{aligned}
, \qquad (2)
$$

where the *j*th claim's number, Claim_j, is also the weight given to that claim's average scores, $\frac{1}{\#\text{scores}_j}(\sum_{k\ \text{in}\ j}\text{WCR}_k)$, in the computation of the total CR score of the TOEIC Writing test.

Constructed-Response (CR) Differential Item Functioning (DIF) Methods

All of the CR DIF methods considered in this study can be described in terms of an average difference in expected and conditional scores of the studied item (*Y*) for the reference group (*G* = *R*, defined as males in this study) and the focal group (*G* = *F*, defined as females in this study) matched across the *m* = 1 to *M* possible scores of a matching variable,

$$
\sum_{m}\left(\frac{n_{m,F}}{N_F}\right)\left[E(Y|\text{Matching}_m, F)\text{-}E(Y|\text{Matching}_m, R)\right], \qquad (3)
$$

where $n_{m,F}$ and N_F denote the conditional and overall sample sizes of the focal group. Equation 3 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,

$$
E(Y|\text{Matching}_m, G)=\mu_{Y|m,G}, \qquad (4)
$$

where $\mu_{Y|m,G}$ denotes the conditional mean of *Y* for the m[th] score of the matching variable in group *G*. The two matching variables used in Equation 4 are the observed total test scores, *CR*, and the total test scores excluding the studied item, *CR - Y* (i.e., excluding the weighted contribution of the studied item to the total TOEIC Speaking test score in Equation 1 or the total TOEIC Writing test score in Equation 2). The *CR* and *CR - Y* designations are used throughout the rest of this report to refer to the two standardized E-DIF approaches.

CR DIF methods based on the PolySIB (Chang et al., 1996; Shealy & Stout, 1993) use expected and conditional *Y* scores that are adjusted and interpreted as conditioned on $T(\text{Matching}_m)$, the estimated true score of the reference and focal groups for the m[th] observed score of the matching variable,

$$E[Y|T_G(\text{Matching}_m)]=$$
$$\mu_{Y|m,G}+\left[\frac{\mu_{Y|m+1,G}-\mu_{Y|m-1,G}}{T_G(\text{Matching}_{m+1})-T_G(\text{Matching}_{m-1})}\right]\left[T(\text{Matching}_m)-T_G(\text{Matching}_m)\right]' \quad (5)$$

where $T_G(\text{Matching}_m)=\mu_{\text{Matching},G}+\text{rel}(\text{Matching}_{.,G})(\text{Matching}_m-\mu_{m,G})$, $\text{rel}(\text{Matching}_{.,G})$ denotes the alpha reliability or internal consistency of the matching variable in group *G* (Kelley, 1923; Shealy & Stout, 1993), and where $T(\text{Matching}_m)=\frac{T_R(\text{Matching}_m)+T_F(\text{Matching}_m)}{2}$.

The two matching variables used in Equation 5 are the estimated true scores of the total test scores, *T*(*CR*), and the estimated true scores of the total test scores excluding the studied item, *T*(*CR* - *Y*) (i.e., excluding the weighted contribution of the studied item to the total TOEIC Speaking test score in Equation 1 or the total TOEIC Writing test score in Equation 2). The *T*(*CR*) and *T*(*CR* - *Y*) designations are used throughout the rest of this report to refer to the two PolySIB approaches.

Prior to computing DIF estimates based on Equations 3–5, the male and female 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). The smoothed frequency data also made it unnecessary to use some data exclusion practices recommended for SIBTEST methods like the PolySIB (e.g., data would not warrant exclusion from calculations of the SIBTEST results when the sample sizes of the reference and focal groups were less than two at any score of the matching variable; Shealy & Stout, 1993, appendix).

Presentation of the Results of the Constructed-Response (CR) Differential Item Functioning (DIF) Method

The presentation of the results of the CR DIF method was based on those used in Moses et al. (in press), in which the results of the various DIF methods were presented as deviations from mean DIF value of all of the methods for each item and score. Interpretations of the DIF results of the CR items can be of interest with respect to the raw score units that reflect the actual score rubrics of the items and also with respect to the standard deviation units of the items, which facilitate comparison of results across items and a uniform set of DIF flagging rules. Because interpretations of both types of results can be useful in practice and of interest to particular readers, two sets of results were produced for the TOEIC Speaking and Writing test items, one set for which the DIF results were in terms of studied items' raw score units and another for which the DIF results were divided by the standard deviations of the studied items from the total examinee data (including female and male examinees). The results of the DIF methods are presented with additional descriptive statistics noted to affect results (including the female-male mean differences divided by the standard deviations for each item and score—i.e., impact) and for the test score excluding *Y* (*CR* – *Y*), the Pearson product moment correlations of the studied item and the *CR* – *Y* scores, and the coefficient alpha reliabilities of the *CR* – *Y* scores. The computation of the correlations and reliabilities involving the *CR* – *Y* scores utilized the item and score weights described in Equations 1 and 2.

Results

The sets of CR DIF results for the TOEIC Speaking test based on raw and standardized units of the studied items are presented in Tables 3 and 4. The sets of CR DIF results for the TOEIC Writing test based on raw and standard deviation units of the studied items are presented in Tables 5 and 6. In these tables, the results of the DIF methods are presented in an order that emphasizes their consistent pattern across the items, reported units (raw or standardized), and the TOEIC Speaking and Writing tests, in which the DIF methods with the most negative to most positive deviations from the mean DIF values were obtained from the *T(CR)*, *T(CR - Y)*, *CR*, and *CR - Y* methods. The pattern of the results of the DIF methods is such that the least extreme results were observed for the most recommended versions of the CR DIF methods, that is, excluding the studied item, *T(CR - Y)*, from the estimated true score of the matching variable for the PolySIB method and including the studied item, *CR*, from the observed score of the matching variable for the standardized E-DIF (Chang et al., 1996; Dorans & Schmitt, 1993). Additional characteristics of the examinee and test data also appeared to affect the DIF results in Tables 3–6, in that the relatively large *F - R* impact values on the matching variables of 0.4–0.6 in combination with the matching variables' relatively low reliabilities of 0.33–0.62 are known to magnify differences between the results of the PolySIB and standardized E-DIF methods.

Table 3

Constructed-Response (CR) Differential Item Functioning (DIF) Results for the TOEIC Speaking Test Form (Raw Score Units of Y)

Item	*F-R* impact on the studied item (*Y*)	*F-R* impact on the matching variable (*CR - Y*)	Correlation (*CR - Y & Y*)	Reliability (*CR - Y*)	Mean DIF value	CR DIF results based on the following matching variables (deviations from the mean DIF value)			
						T(CR)	*T(CR - Y)*	*CR*	*CR - Y*
SCR1: Read a text aloud (pronunciation)	0.44	0.40	0.40	0.62	0.12	-0.03	-0.02	0.02	0.03
SCR1: Read a text aloud (intonation)	0.48	0.40	0.38	0.62	0.15	-0.03	-0.02	0.02	0.02
SCR2: Read a text aloud (pronunciation)	0.20	0.41	0.39	0.62	-0.01	-0.03	-0.02	0.02	0.02
SCR2: Read a text aloud (intonation)	0.25	0.41	0.46	0.62	0.00	-0.03	-0.03	0.02	0.03
SCR3: Describe a picture	0.36	0.41	0.48	0.62	0.04	-0.04	-0.04	0.04	0.04
SCR4: Respond to questions	0.12	0.41	0.24	0.62	-0.06	-0.06	0.00	0.02	0.05
SCR5: Respond to questions	0.44	0.39	0.38	0.61	0.13	-0.05	-0.02	0.02	0.04
SCR6: Respond to questions	0.16	0.43	0.39	0.60	-0.04	-0.05	-0.03	0.04	0.05
SCR7: Respond to questions using information provided	0.09	0.41	0.24	0.63	-0.02	-0.03	-0.01	0.01	0.02
SCR8: Respond to questions using information provided	0.29	0.40	0.33	0.62	0.05	-0.05	-0.02	0.02	0.04
SCR9: Respond to questions using information provided	0.18	0.39	0.29	0.62	0.01	-0.03	-0.01	0.01	0.03
SCR10: Propose a solution	0.23	0.44	0.52	0.57	-0.08	-0.07	-0.03	0.03	0.07
SCR11: Express an opinion	0.29	0.43	0.52	0.58	-0.03	-0.08	-0.03	0.03	0.07

Note. *CR* = observed total test scores, *CR - Y* = total test scores excluding the studied item, *F* = focal group, *R* = reference group, SCR = TOEIC Speaking test constructed-response item, *T*(*CR*) = estimated true scores of the total test scores, *T*(*CR - Y*) = estimated true scores of the total test scores excluding the studied item.

Table 4

Constructed-Response (CR) Differential Item Functioning (DIF) Results for the TOEIC Speaking Test Form (Standard Deviation Units of Y)

Item	*F-R* impact on the studied item (*Y*)	*F-R* impact on the matching variable (*CR - Y*)	Correlation (*CR - Y & Y*)	Reliability (*CR - Y*)	Mean DIF value	CR DIF results based on the following matching variables (deviations from the mean DIF value)			
						T(CR)	*T(CR - Y)*	*CR*	*CR - Y*
SCR1: Read a text aloud (pronunciation)	0.44	0.40	0.40	0.62	0.22	-0.06	-0.03	0.03	0.05
SCR1: Read a text aloud (intonation)	0.48	0.40	0.38	0.62	0.28	-0.04	-0.04	0.03	0.04
SCR2: Read a text aloud (pronunciation)	0.20	0.41	0.39	0.62	-0.03	-0.05	-0.03	0.04	0.05
SCR2: Read a text aloud (intonation)	0.25	0.41	0.46	0.62	0.00	-0.05	-0.05	0.05	0.05
SCR3: Describe a picture	0.36	0.41	0.48	0.62	0.07	-0.06	-0.06	0.06	0.06
SCR4: Respond to questions	0.12	0.41	0.24	0.62	-0.07	-0.07	0.00	0.02	0.06
SCR5: Respond to questions	0.44	0.39	0.38	0.61	0.19	-0.07	-0.03	0.03	0.07
SCR6: Respond to questions	0.16	0.43	0.39	0.60	-0.05	-0.07	-0.05	0.05	0.07
SCR7: Respond to questions using information provided	0.09	0.41	0.24	0.63	-0.05	-0.06	-0.02	0.03	0.05
SCR8: Respond to questions using information provided	0.29	0.40	0.33	0.62	0.07	-0.07	-0.02	0.03	0.06
SCR9: Respond to questions using information provided	0.18	0.39	0.29	0.62	0.01	-0.06	-0.01	0.02	0.05
SCR10: Propose a solution	0.23	0.44	0.52	0.57	-0.13	-0.11	-0.05	0.05	0.11
SCR11: Express an opinion	0.29	0.43	0.52	0.58	-0.05	-0.12	-0.04	0.04	0.11

Note. CR = observed total test scores, *CR - Y* = total test scores excluding the studied item, *F* = focal group, *R* = reference group, SCR = TOEIC Speaking test constructed-response item, *T*(*CR*) = estimated true scores of the total test scores, *T*(*CR - Y*) = estimated true scores of the total test scores excluding the studied item.

Table 5

Constructed-Response (CR) Differential Item Functioning (DIF) Results for the TOEIC Writing Test Form (Raw Score Units of Y)

Item	*F-R* impact on the studied item (*Y*)	*F-R* impact on the matching variable (*CR - Y*)	Correlation (*CR - Y & Y*)	Reliability (*CR - Y*)	Mean DIF value	CR DIF results based on the following matching variables (deviations from the mean DIF value)			
						T(CR)	*T(CR - Y)*	*CR*	*CR - Y*
WCR1: Write a sentence based on a picture	0.18	0.52	0.16	0.44	0.02	-0.04	-0.01	0.02	0.03
WCR2: Write a sentence based on a picture	0.26	0.52	0.18	0.44	0.03	-0.06	-0.01	0.02	0.04
WCR3: Write a sentence based on a picture	0.20	0.51	0.23	0.43	-0.02	-0.09	-0.02	0.04	0.07
WCR4: Write a sentence based on a picture	0.32	0.51	0.35	0.42	0.01	-0.12	-0.05	0.07	0.10
WCR5: Write a sentence based on a picture	0.31	0.51	0.28	0.43	0.04	-0.11	-0.03	0.05	0.09
WCR6: Respond to a written request	0.40	0.50	0.38	0.33	-0.04	-0.18	-0.08	0.08	0.17
WCR7: Respond to a written request	0.50	0.47	0.37	0.31	0.03	-0.23	-0.07	0.09	0.21
WCR8: Write an opinion essay	0.31	0.63	0.39	0.43	-0.12	-0.24	0.02	0.08	0.15

Note. CR = observed total test scores, *CR - Y* = total test scores excluding the studied item, *F* = focal group, *R* = reference group, *T*(*CR*) = estimated true scores of the total test scores, *T*(*CR - Y*) = estimated true scores of the total test scores excluding the studied item, WCR = TOEIC Writing test constructed-response item.

Table 6

Constructed-Response (CR) Differential Item Functioning (DIF) Results for the TOEIC Writing Test Form (Standard Deviation Units of Y)

Item	*F-R* impact on the studied item (*Y*)	*F-R* impact on the matching variable (*CR - Y*)	Correlation (*CR - Y* & *Y*)	Reliability (*CR - Y*)	Mean DIF value	CR DIF results based on the following matching variables (deviations from the mean DIF value)			
						T(CR)	*T(CR - Y)*	*CR*	*CR - Y*
WCR1: Write a sentence based on a picture	0.18	0.52	0.16	0.44	0.05	-0.10	-0.03	0.05	0.08
WCR2: Write a sentence based on a picture	0.26	0.52	0.18	0.44	0.06	-0.12	-0.02	0.04	0.09
WCR3: Write a sentence based on a picture	0.20	0.51	0.23	0.43	-0.02	-0.15	-0.03	0.06	0.11
WCR4: Write a sentence based on a picture	0.32	0.51	0.35	0.42	0.02	-0.19	-0.08	0.11	0.16
WCR5: Write a sentence based on a picture	0.31	0.51	0.28	0.43	0.06	-0.16	-0.04	0.08	0.13
WCR6: Respond to a written request	0.40	0.50	0.38	0.33	-0.07	-0.28	-0.12	0.14	0.26
WCR7: Respond to a written request	0.50	0.47	0.37	0.31	0.03	-0.30	-0.08	0.11	0.28
WCR8: Write an opinion essay	0.31	0.63	0.39	0.43	-0.22	-0.45	0.03	0.15	0.28

Note. CR = observed total test scores, *CR - Y* = total test scores excluding the studied item, *F* = focal group, *R* = reference group, *T*(*CR*) = estimated true scores of the total test scores, *T*(*CR - Y*) = estimated true scores of the total test scores excluding the studied item, WCR = TOEIC Writing test constructed-response item.

Although the negative-to-positive results pattern was consistent across the items and tests, it was more visible in the items at the end of each test that had greater ranges of raw scores (Tables 1 and 2) and made larger weighted contributions to the total test scores (Equations 1 and 2). To assess the results patterns observed in Tables 3–6 in greater detail, the female and male conditional means used in the calculations for the *T*(*CR - Y*), *T*(*CR*), *CR*, and *CR - Y* methods (Equations 3–5) were plotted for two items of the TOEIC Writing test: WCR1 (which had a relatively narrow possible score range and made a relatively small weighted contribution to the total score of the TOEIC Writing test; see Figure 1) and WCR8 (which had a relatively wide possible score range and made a relatively large weighted contribution to the total score of the TOEIC Writing test; see Figure 2). Because the *T*(*CR - Y*), *T*(*CR*), *CR*, and *CR - Y* methods weight the conditional mean differences of the male and female examinees by the distribution of the female group on the matching variable (Equation 1), the rescaled female distribution is also plotted in the figures. These figures, which are particularly useful for revealing the effects of including a highly weighted studied item in the matching variable, show studied items that account for large portions of the total test score can result in series of conditional means that are essentially step functions that exhibit little variability when the matching variable is used in its observed form (e.g., the *CR* results in Figure 2) and little variability that is shifted in favor of the underperforming males when the matching variable is used in its estimated true score form (e.g., the *T*(*CR*) results in Figure 2). Although Figures 1 and 2 indicate that the systematic irregularities are more visibly apparent for item WCR8 than for item WCR1, the consistent results patterns in Tables 3–6 suggest that similar issues affect the DIF methods across all the test items.

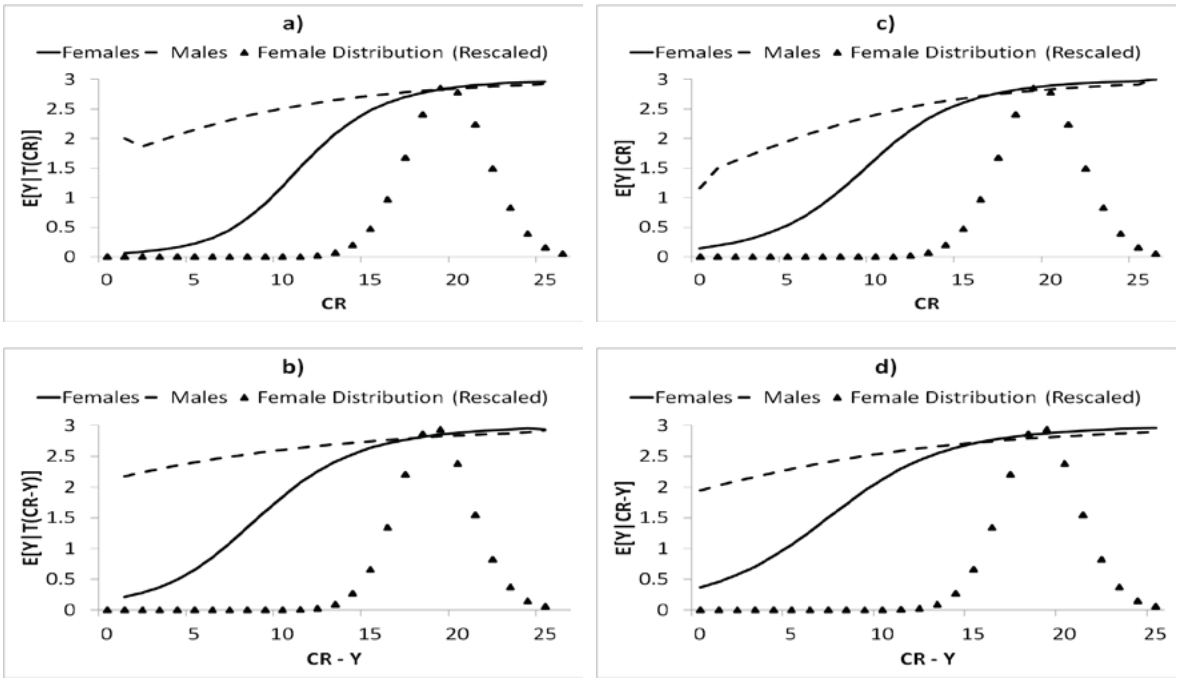

***Figure 1*. For TOEIC Writing test constructed-response item WCR1, the conditional expected item scores for the female and male groups and the female group's rescaled matching variable distribution corresponding to the matching variables of the TOEIC Writing test: (a) *T*(*CR - Y*) (estimated true scores of the total test scores excluding the studied item), (b) *T*(*CR*) (estimated true scores of the total test scores), (c) *CR* (observed total test scores), and (d) *CR - Y* (total test scores excluding the studied item).**

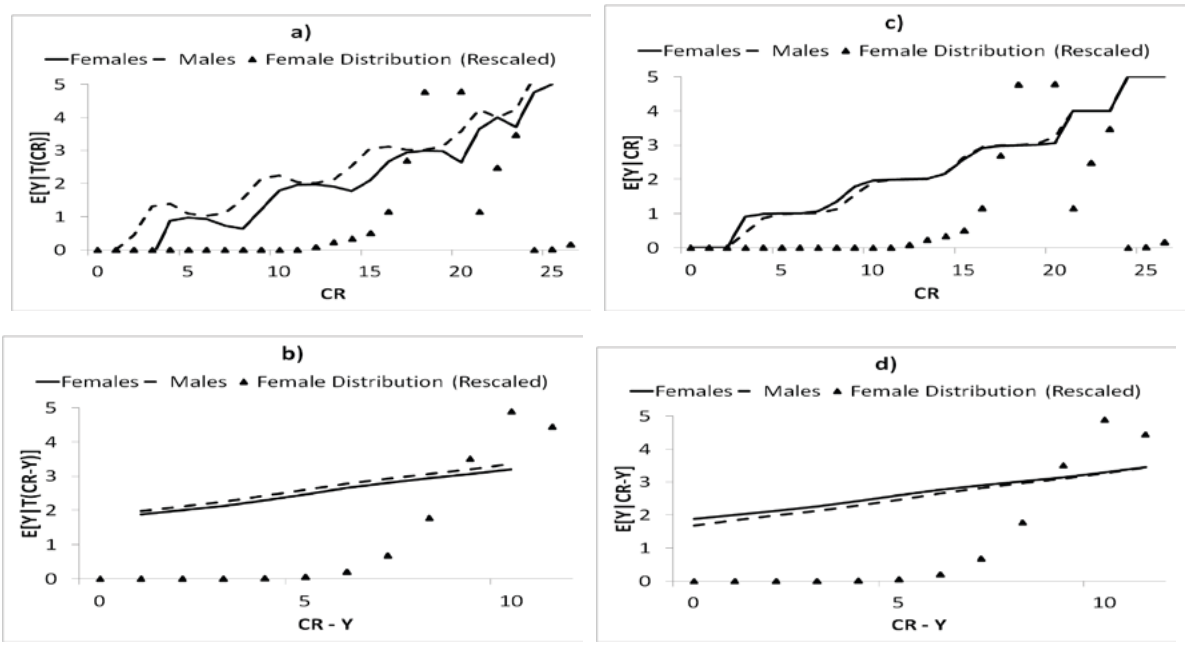

***Figure 2*. For TOEIC Writing test constructed-response item WCR8, the conditional expected item scores for the female and male groups and the female group's rescaled matching variable distribution corresponding the matching variables of the TOEIC Writing test: (a) *T*(*CR - Y*) (estimated true scores of the total test scores excluding the studied item), (b) *T*(*CR*) (estimated true scores of the total test scores), (c) *CR* (observed total test scores), and (d) *CR - Y* (total test scores excluding the studied item).**

An attempt was made to address the systematic irregularities observed in the DIF method results (Figure 2) as well as the short and relatively unreliable matching variables available from the TOEIC Speaking and Writing tests. By noting that both tests were taken by a single examinee group, that the correlation of the total scores of the two tests was not exceedingly low (0.64), and that the total scores of the TOEIC Speaking and Writing tests had similar ranges (0–26 and 0–24), the results of the CR DIF methods were reproduced using the combined TOEIC Speaking and Writing test scores as the matching variables. Additional sets of results for all TOEIC Speaking and Writing items in the raw score units and standard deviation units of the items are presented in Tables 7 and 8.

Although the results in Tables 7 and 8 exhibit negative-to-positive orderings of the DIF method values similar to those in Tables 3–6, the ranges of the methods' mean deviations are narrower in Tables 7 and 8. Tables 7 and 8 also show that matching variables based on the combined TOEIC Speaking and Writing total scores are more reliable (i.e., 0.67–0.72 for combined TOEIC Speaking and Writing total scores vs. 0.58–0.63 for TOEIC Speaking scores only and 0.31–0.44 for TOEIC Writing scores only) and usually more highly correlated with the studied items (i.e., 0.16–0.59 for combined TOEIC Speaking and Writing total scores vs. 0.24–0.52 for TOEIC Speaking scores only and 0.16–0.39 for TOEIC Writing scores only). The smaller gains in correlations than the gains in reliabilities in

Tables 7 and 8 were possibly due to the narrow score ranges of the studied items. Comparisons of the CR DIF results for the TOEIC Speaking and Writing items in Tables 7 and 8 with the results in Tables 3–6 indicate that the narrow ranges of mean deviations, increased reliabilities, and increased correlations of Tables 7 and 8 are more apparent for the TOEIC Writing items than for the TOEIC Speaking items. The greater differences for the TOEIC Writing items correspond to the TOEIC Writing test being shorter and less reliable than the TOEIC Speaking test, so that the use of the combined TOEIC Speaking and Writing score as a CR DIF matching variable had a greater impact on the CR DIF results and reliabilities for the TOEIC Writing items than for the TOEIC Speaking items. Figures 3 and 4 present the series of conditional means of TOEIC Writing items WCR1 and WCR8 based on the combined TOEIC Speaking and Writing scores, showing less of the systematic irregularities due to including highly weighted studied items in the DIF matching variable (Figure 4 vs. Figure 2).

Discussion

The CR DIF methods used in this study to evaluate TOEIC Speaking and Writing items were selected based on their use in evaluating CR DIF in mixed format tests (Moses et al., in press). Conducting exploratory evaluations focusing on the comparison of several CR DIF methods was useful for addressing the uncertainties about CR DIF evaluations expressed in informal surveys of ETS statistical staff and also for reflecting the prominence of these methods in CR DIF research (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 general opinion from prior research and practice for CR DIF in mixed format tests is that DIF results obtained from matching variables based on all items of a test, both MC and CR, are preferable to the DIF results based on matching variables with only MC or CR scores.

Somewhat similar to general opinions about matching variables for CR DIF evaluations on mixed format tests, the current study suggested that CR DIF results from using a combination of the scores from TOEIC Speaking and Writing tests as the matching variable were improved relative to the results obtained from using only one of the two test scores as the matching variable. The observed improvements were (usually) higher correlations with the studied items, higher reliabilities, and greater consistency in the DIF method results (especially for the TOEIC Writing test). Additional studies would be useful for considering how these results compare to those obtained from other TOEIC test forms, different approaches to combining matching variables from separate tests (e.g., bivariate matching in addition to summed scores), and perhaps using additional scores, such as from the TOEIC Listening and Reading tests in a combined CR DIF matching variable.

Table 7

Constructed-Response (CR) Differential Item Functioning (DIF) Results for the Combined TOEIC Speaking and TOEIC Writing Test Forms Using the Combined TOEIC Speaking and Writing Score as the CR Matching Variable (Raw Score Units of Y)

Item	*F-R* impact on the studied item (*Y*)	*F-R* impact on the matching variable (*CR - Y*)	Correlation (*CR - Y & Y*)	Reliability (*CR - Y*)	Mean DIF value	CR DIF results based on the following matching variables (deviations from the mean DIF value)			
						T(CR)	*T(CR - Y)*	*CR*	*CR - Y*
SCR1: Read a text aloud (pronunciation)	0.44	0.53	0.42	0.72	0.09	-0.02	-0.02	0.02	0.02
SCR1: Read a text aloud (intonation)	0.48	0.54	0.39	0.72	0.12	-0.02	-0.01	0.02	0.02
SCR2: Read a text aloud (pronunciation)	0.20	0.53	0.36	0.72	-0.04	-0.03	-0.02	0.02	0.03
SCR2: Read a text aloud (intonation)	0.25	0.54	0.43	0.72	-0.01	-0.02	-0.02	0.02	0.02
SCR3: Describe a picture	0.36	0.53	0.49	0.71	0.01	-0.03	-0.03	0.03	0.03
SCR4: Respond to questions	0.12	0.53	0.26	0.71	-0.07	-0.04	-0.01	0.01	0.03
SCR5: Respond to questions	0.44	0.52	0.42	0.71	0.09	-0.03	-0.02	0.02	0.04
SCR6: Respond to questions	0.16	0.53	0.39	0.71	-0.07	-0.04	-0.02	0.03	0.04
SCR7: Respond to questions using information provided	0.09	0.53	0.25	0.72	-0.03	-0.02	-0.01	0.01	0.02
SCR8: Respond to questions using information provided	0.29	0.52	0.38	0.71	0.02	-0.03	-0.02	0.02	0.03
SCR9: Respond to questions using information provided	0.18	0.53	0.31	0.72	-0.01	-0.02	-0.01	0.01	0.02
SCR10: Propose a solution	0.23	0.56	0.59	0.67	-0.13	-0.05	-0.04	0.04	0.06
SCR11: Express an opinion	0.29	0.53	0.53	0.68	-0.05	-0.06	-0.02	0.02	0.06
WCR1: Write a sentence based on a picture	0.18	0.54	0.16	0.72	0.04	-0.01	-0.01	0.00	0.01
WCR2: Write a sentence based on a picture	0.26	0.52	0.16	0.72	0.06	-0.01	0.00	0.01	0.01
WCR3: Write a sentence based on a picture	0.20	0.54	0.27	0.72	0.02	-0.02	-0.02	0.02	0.02
WCR4: Write a sentence based on a picture	0.32	0.53	0.38	0.71	0.06	-0.03	-0.03	0.03	0.03
WCR5: Write a sentence based on a picture	0.31	0.53	0.28	0.72	0.10	-0.02	-0.01	0.02	0.02
WCR6: Respond to a written request	0.40	0.49	0.42	0.70	0.06	-0.07	0.00	0.01	0.06
WCR7: Respond to a written request	0.50	0.47	0.43	0.69	0.14	-0.08	0.01	0.00	0.07
WCR8: Write an opinion essay	0.31	0.54	0.57	0.71	-0.05	-0.05	-0.01	0.01	0.05

Note. *CR* = observed total test scores, *CR - Y* = total test scores excluding the studied item, *F* = focal group, *R* = reference group, SCR = TOEIC Speaking test constructed-response item, *T*(*CR*) = estimated true scores of the total test scores, *T*(*CR - Y*) = estimated true scores of the total test scores excluding the studied item, WCR = TOEIC Writing test constructed-response item.

Table 8

Constructed-Response (CR) Differential Item Functioning (DIF) Results for the Combined TOEIC Speaking and TOEIC Writing Test Forms Using the Combined TOEIC Speaking and Writing Score as the CR Matching Variable (Standard Deviation Units of Y)

Item	*F-R* impact on the studied item (*Y*)	*F-R* impact on the matching variable (*CR - Y*)	Correlation (*CR - Y & Y*)	Reliability (*CR - Y*)	Mean DIF value	CR DIF results based on the following matching variables (deviations from the mean DIF value)			
						T(CR)	*T(CR - Y)*	*CR*	*CR - Y*
SCR1: Read a text aloud (pronunciation)	0.44	0.53	0.42	0.72	0.18	-0.04	-0.04	0.04	0.04
SCR1: Read a text aloud (intonation)	0.48	0.54	0.39	0.72	0.24	-0.04	-0.04	0.04	0.04
SCR2: Read a text aloud (pronunciation)	0.20	0.53	0.36	0.72	-0.07	-0.04	-0.03	0.03	0.04
SCR2: Read a text aloud (intonation)	0.25	0.54	0.43	0.72	-0.02	-0.04	-0.05	0.04	0.04
SCR3: Describe a picture	0.36	0.53	0.49	0.71	0.02	-0.05	-0.05	0.05	0.05
SCR4: Respond to questions	0.12	0.53	0.26	0.71	-0.09	-0.04	-0.01	0.02	0.04
SCR5: Respond to questions	0.44	0.52	0.42	0.71	0.14	-0.06	-0.03	0.03	0.05
SCR6: Respond to questions	0.16	0.53	0.39	0.71	-0.08	-0.06	-0.03	0.03	0.05
SCR7: Respond to questions using information provided	0.09	0.53	0.25	0.72	-0.06	-0.04	-0.02	0.02	0.04
SCR8: Respond to questions using information provided	0.29	0.52	0.38	0.71	0.03	-0.05	-0.03	0.03	0.04
SCR9: Respond to questions using information provided	0.18	0.53	0.31	0.72	-0.02	-0.04	-0.02	0.03	0.04
SCR10: Propose a solution	0.23	0.56	0.59	0.67	-0.21	-0.08	-0.07	0.06	0.09
SCR11: Express an opinion	0.29	0.53	0.53	0.68	-0.08	-0.09	-0.03	0.04	0.09
WCR1: Write a sentence based on a picture	0.18	0.54	0.16	0.72	0.09	-0.02	-0.01	0.01	0.02
WCR2: Write a sentence based on a picture	0.26	0.52	0.16	0.72	0.12	-0.02	-0.01	0.01	0.02
WCR3: Write a sentence based on a picture	0.20	0.54	0.27	0.72	0.02	-0.03	-0.02	0.03	0.03
WCR4: Write a sentence based on a picture	0.32	0.53	0.38	0.71	0.09	-0.05	-0.04	0.04	0.05
WCR5: Write a sentence based on a picture	0.31	0.53	0.28	0.72	0.15	-0.04	-0.02	0.02	0.03
WCR6: Respond to a written request	0.40	0.49	0.42	0.70	0.09	-0.10	0.00	0.01	0.09
WCR7: Respond to a written request	0.50	0.47	0.43	0.69	0.19	-0.11	0.01	0.00	0.09
WCR8: Write an opinion essay	0.31	0.54	0.57	0.71	-0.09	-0.10	-0.02	0.02	0.09

Note. *CR* = observed total test scores, *CR - Y* = total test scores excluding the studied item, *F* = focal group, *R* = reference group, SCR = TOEIC Speaking test constructed-response item, *T*(*CR*) = estimated true scores of the total test scores, *T*(*CR - Y*) = estimated true scores of the total test scores excluding the studied item, WCR = TOEIC Writing test constructed-response item.

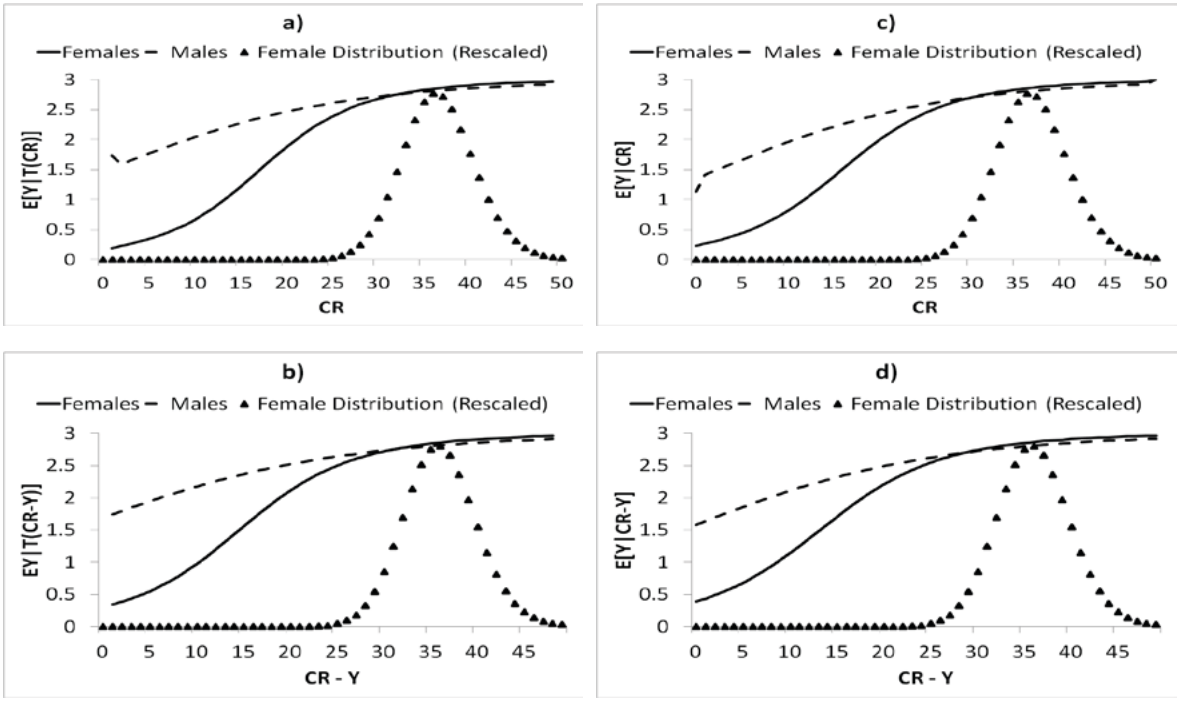

***Figure 3*. For TOEIC Writing test constructed-response item WCR1, the conditional expected item scores for the female and male groups and the female group's rescaled matching variable distribution corresponding to the matching variables of the TOEIC Writing test: (a) *T*(*CR - Y*) (estimated true scores of the total test scores excluding the studied item), (b) *T*(*CR*) (estimated true scores of the total test scores), (c) *CR* (observed total test scores), and (d) *CR - Y* (total test scores excluding the studied item).**

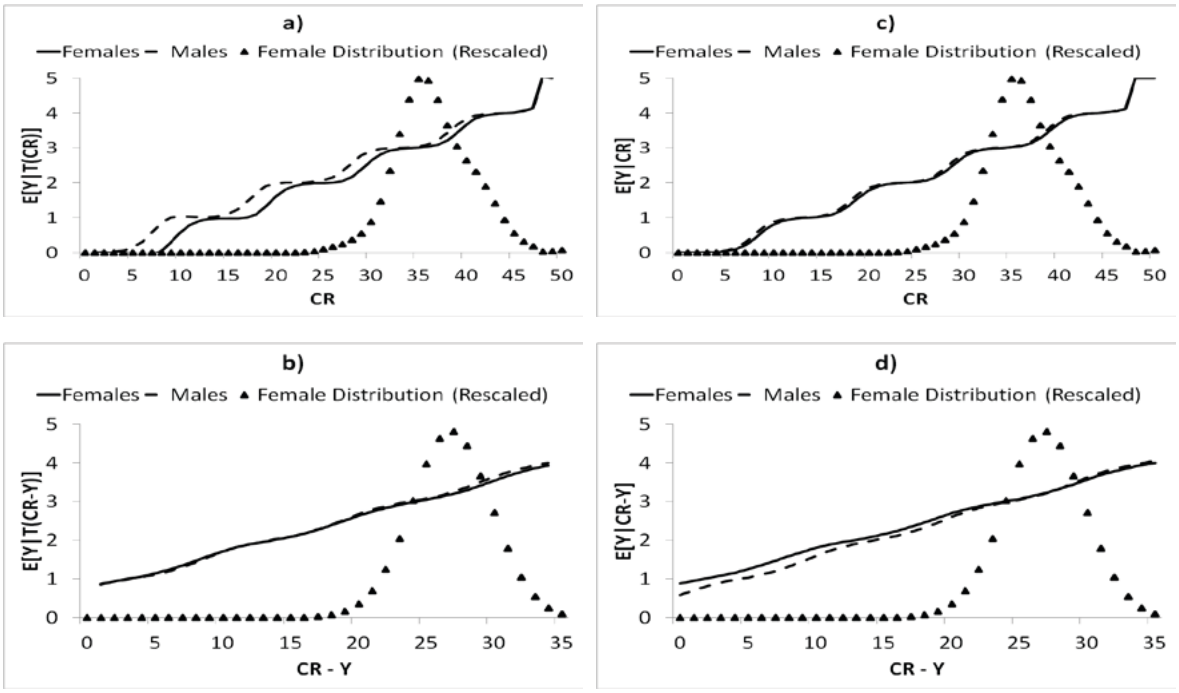

Figure 4. For TOEIC Writing test constructed-response item WCR8, the conditional expected item scores for the female and male groups and the female group's rescaled matching variable distribution corresponding to the matching variables of the TOEIC Writing test (a) *T*(*CR - Y*) (estimated true scores of the total test scores excluding the studied item), (b) *T*(*CR*) (estimated true scores of the total test scores), (c) *CR* (observed total test scores), and (d) *CR - Y* (total test scores excluding the studied item).

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., & 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 Associates.

Educational Testing Service. (2010). *TOEIC Speaking & Writing user guide*. Princeton, NJ: Author.

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., Liu, J., Tan, X., Deng, W., & Dorans, N. J. (in press). *Constructed-response DIF evaluations for mixed format tests* (Research Report). Princeton, NJ: Educational Testing Service.

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.

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.