Article

Applied Psychological Measurement
34(3) 151–165

DOI: 10.1177/0146621609359284
http://apm.sagepub.com

Modeling DIF Effects Using Distractor-Level Invariance Effects: Implications for Understanding the Causes of DIF

Randall D. Penfield[1]

Abstract

In 2008, Penfield showed that measurement invariance across all response options of a multiple-choice item (correct option and the *J* distractors) can be modeled using a nominal response model that included a differential distractor functioning (DDF) effect for each of the *J* distractors. This article extends this concept to consider how the differential item functioning (DIF) effect (i.e., the conditional between-group differences in the probability of correct response) is determined by the *J* DDF effects. In particular, this article shows how the DIF effect can be modeled as a function of the *J* DDF effects and thus reveals the conditions that must hold for uniform DIF, nonuniform DIF, and crossing DIF to exist. The results provide insight into the potential item-level properties that lead to uniform, nonuniform, and crossing DIF. The findings may shed light on the etiology of different forms of DIF, which may help analysts target the particular causes of the DIF effect.

Keywords

differential item functioning, differential distractor functioning, multiple-choice items

Measurement invariance at the item level is defined as the independence of group membership and item response conditional on the intended target trait (Millsap, 2005; Millsap & Meredith, 1992). The condition of invariance is desirable because it ensures that the responses to the item reflect only content-relevant variance. Violation of invariance suggests the presence of an unintended item-level multidimensionality (Ackerman, 1992; Camilli, 1992) that may be caused by a biasing factor in the item (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Invariance is important because it concerns whether the scores have the same meaning for two or more groups.

When scaling is conducted using multiple-choice item formats, it is customary to dichotomize the item response into two categories: one category representing the correct option and one

[1]University of Miami, Coral Gables, Florida

Corresponding Author:

Randall D. Penfield, University of Miami, P.O. Box 248065, Coral Gables, FL 33124-2040
Email: penfield@miami.edu

category representing all distractor options. As a result, the study of measurement invariance in multiple-choice items typically considers the independence of the dichotomized response variable and group membership conditional on target trait. A violation of this form of invariance in dichotomously scored items is commonly referred to as differential item functioning (DIF; Holland & Thayer, 1988). Because widely used response models for multiple-choice items are based on modeling the probability of correct response as a function of the target trait, DIF in multiple-choice items is commonly conceptualized as a between-group difference in the conditional probability of correct response.

When interpreting the meaning of the DIF observed in a dichotomously scored item, the form of the DIF effect becomes important because different forms of DIF yield differential evidence concerning the potential causes of the DIF effect. Two general forms of DIF, known as uniform DIF and nonuniform DIF, receive widespread use in the DIF literature (Camilli & Shepard, 1994). The DIF effect is deemed uniform when the ratio of the odds of correct response for the two groups is independent of the target trait (Hanson, 1998). It is relevant to note that the condition of uniform DIF corresponds to the condition of a between-group difference in the difficulty parameter of the one- and two-parameter logistic models (Lord, 1980) when the discrimination parameter is constant across groups. The DIF effect is deemed nonuniform when the ratio of the odds of correct response for the two groups varies as a function of the target trait. More specific forms of nonuniform DIF have also been described in the literature (Finch & French, 2006; Hanson, 1998; Penfield, 2003) and include unidirectional DIF (i.e., when the DIF effect is not uniform, but the same group experiences a relative advantage across all levels of the target trait) and crossing DIF (i.e., when the relative advantage favors different groups at different ranges of the target trait continuum). In practice, crossing DIF is a particularly troublesome form of DIF because of the insensitivity of several widely used DIF methods (i.e., proportion-difference and common odds ratio approaches; Penfield & Camilli, 2007) to detect this form of DIF. Given the importance of crossing DIF to practical considerations of DIF detection methodology, this article will often distinguish between nonuniform DIF (the variability of the conditional odds ratio) and crossing DIF (that portion of the nonuniform DIF effect that cancels because of differences in sign).

Although the presence of DIF in a multiple-choice item indicates a violation of invariance, the results of a DIF analysis alone do not provide information as to where among the response options the DIF effect is being manifested. Studying where among the response options the invariance is occurring can provide valuable information concerning the causes of the DIF effect. To address this issue, researchers have proposed the framework of differential distractor functioning (DDF; Green, Crone, & Folk, 1989; Penfield, 2008; Schmitt & Dorans, 1990), which is aimed at evaluating the invariance effect associated with each of the J distractors of the item. The pattern of J DDF effects can shed light on the cause of DIF. The presence of a consistent DDF effect across all distractors provides evidence that the factor causing the DIF may reside in the item stem or the correct option, whereas DDF, present primarily in isolated distractors, provides evidence that the causal property may be related to properties of particular distractors (Penfield, 2008).

Not only can DDF help to provide information about which response options are responsible for the DIF effect, DDF can also serve as a framework for understanding how DIF effects are generated as a result of between-group differences in distractor-level properties. Specifically, Penfield (2008) showed that under the nominal response model (NRM; Bock, 1972) the violation of invariance in a multiple-choice item can be modeled using the J DDF effects. Because DIF corresponds to a specific form of invariance violation, it follows that the DIF effect can be expressed as a function of the DDF effects when the data follow the NRM, and the resulting functional form can be used to determine what DDF patterns lead to different forms of DIF. Understanding how DDF effects relate to the overall DIF effect can then help us gain clarity on the underlying item properties that lead to different forms of DIF.

To date, the specific functional form relating the DIF effect to the item's DDF effects has not been presented in the literature. This article serves to address this issue. To this end, the remainder of this article aims to demonstrate (a) how the DIF effect, modeled under the NRM, can be expressed as a function of the J DDF effects; (b) what forms of DDF patterns are required for different forms of DIF to hold (i.e., uniform, nonuniform, crossing); (c) how the relationship between DIF and DDF effects can be used to improve our understanding of DIF effects in multiple-choice items; and (d) how DIF and DDF analyses can be used in concert to most effectively evaluate the presence and causes of a DIF effect.

Modeling Distractor-Level Invariance Effects Under the NRM

Consider a multiple-choice item containing options $j = 1, 2, \ldots, m$. Let the option $j = m$ be reserved for the correct option, and the options $j = 1, 2, \ldots, J$, where $J = m-1$, be reserved for the distractor options. The NRM can be used to model the conditional probability associated with the correct option and each of the J distractor options. Under the NRM, the probability of selecting the correct option conditional on target trait (θ) can be parameterized according to

$$P(Y = m|\theta) = \frac{1}{1 + \sum_{j=1}^{J} \exp(-c_j - a_j\theta)} \tag{1}$$

and the probability of selecting the jth distractor conditional on target trait can be parameterized according to

$$P(Y = j|\theta) = \frac{\exp(-c_j - a_j\theta)}{1 + \sum_{j=1}^{J} \exp(-c_j - a_j\theta)}. \tag{2}$$

The formulation of the NRM presented in Equations 1 and 2 is a simple transformation of the original parameterization presented by Bock (1972). The parameterization shown in Equations 1 and 2 is consistent with that previously described by Thissen, Steinberg, and Fitzpatrick (1989, p. 165) and is used here because it portrays the item parameters in a form that will prove useful when a DDF effect parameter is included in the model (as described below). The relationship between Bock's original form and the form shown in Equations 1 and 2 can be described as follows: (a) let $\tilde{c}_m$ and $\tilde{a}_m$ denote the location and slope parameters of the reference category used under Bock's (1972) formulation, (b) let $\tilde{c}_j$ and $\tilde{a}_j$ denote the location and slope parameters of the jth alternative category under Bock's (1972) formulation, and (c) the c_j and a_j parameters of Equations 1 and 2 can be obtained by $c_j = \tilde{c}_m - \tilde{c}_j$ and $a_j = \tilde{a}_m - \tilde{a}_j$.

The NRM can be extended to consider distractor-level invariance effects (ω_j) such that the probability of selecting the correct option conditional on θ is given by

$$P(Y = m|\theta) = \frac{1}{1 + \sum_{j=1}^{J} \exp(-c_j - a_j\theta - G\omega_j)} \tag{3}$$

and the probability of selecting the jth distractor conditional on θ is given by

$$P(Y = j|\theta) = \frac{\exp(-c_j - a_j\theta - G\omega_j)}{1 + \sum_{j=1}^{J} \exp(-c_j - a_j\theta - G\omega_j)}, \tag{4}$$

where G corresponds to a grouping variable such that $G = 0$ for the focal (F) group and $G = 1$ for the reference (R) group. A positive value of ω_j indicates that the correct option is relatively more attractive than the jth distractor for the reference group than for the focal group, and a negative value of ω_j indicates that the correct option is relatively more attractive than the jth distractor for the focal group than for the reference group. As described by Penfield (2008), the parameterization shown in Equations 3 and 4 models all item-level invariance using the J DDF effects. In interpreting the value of ω_j, it is important to recognize that a nonzero value of ω_j reflects a disproportionate selection of the jth distractor in relation to the correct option, and thus the mechanism underlying a nonzero ω_j can reflect a disproportionate attractiveness of either the jth distractor or the correct option. That is, a distractor being more frequently selected could be a result of either a disproportionately higher attractiveness of the distractor or a disproportionately lower attractiveness of the correct option.

Expressing the DIF Effect as a Function of the DDF Effects

The DIF effect pertains to the between-group difference in the probability of correct response conditional on θ. The magnitude and form of the DIF effect is determined by the nature of the conditional between-group difference in the probability of correct response, given by

$$P(Y = m|\theta, G = 1) \neq P(Y = m|\theta, G = 0). \tag{5}$$

Several approaches can be used to quantify the magnitude of the conditional DIF effect described in Equation 5. One approach is to consider the simple difference between reference and focal group probabilities conditional on θ (Raju, 1988). Although this difference has intuitive appeal (i.e., a difference of 0.1 is easily interpreted), it has an unappealing property of varying greatly across the θ continuum. As a result, it is often useful to quantify the conditional DIF effect using the natural logarithm of the ratio of the conditional odds of correct response (Hanson, 1998), such that the conditional log-odds ratio is expressed as

$$\Lambda(\theta) = \ln\left[\frac{P(Y = m|\theta, G = 1)/[1 - P(Y = m|\theta, G = 1)]}{P(Y = m|\theta, G = 0)/[1 - P(Y = m|\theta, G = 0)]}\right]. \tag{6}$$

The condition of $\Lambda(\theta) = 0$ corresponds to the absence of a between-group difference in the conditional probability of correct response. The utility of using the conditional log-odds ratio resides in the link between the odds ratio and uniform and nonuniform DIF effects, as detailed in the following sections. In this article, the conditional DIF effect is defined using $\Lambda(\theta)$ as specified by Equation 6.

To describe how the DIF effect can be modeled using the J DDF effects, let us consider the parameterization of the NRM given by Equations 3 and 4. Substituting this parameterization into Equation 6 yields the following form of $\Lambda(\theta)$

$$\begin{aligned}
\Lambda(\theta) &= \ln\left\{\left[\frac{\frac{1}{1+\sum_{j=1}^{J}\exp(-c_j-a_j\theta-\omega_j)}}{\frac{\sum_{j=1}^{J}\exp(-c_j-a_j\theta-\omega_j)}{1+\sum_{j=1}^{J}\exp(-c_j-a_j\theta-\omega_j)}}\right] \times \left[\frac{\frac{1}{1+\sum_{j=1}^{J}\exp(-c_j-a_j\theta)}}{\frac{\sum_{j=1}^{J}\exp(-c_j-a_j\theta)}{1+\sum_{j=1}^{J}\exp(-c_j-a_j\theta)}}\right]^{-1}\right\} \\
&= \ln\left[\frac{\sum_{j=1}^{J}\exp(-c_j - a_j\theta)}{\sum_{j=1}^{J}\exp(-c_j - a_j\theta - \omega_j)}\right] \\
&= \ln\left[\frac{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_j\theta)}{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_j\theta)\exp(-\omega_j)}\right].
\end{aligned} \tag{7}$$

Note that Equation 7 expresses the conditional DIF effect, $\Lambda(\theta)$, as a function of c_j, a_j, θ, and ω_j. Having expressed $\Lambda(\theta)$ as a function of the J DDF effects, we can now turn our attention to identifying the specific conditions required in order for uniform DIF, nonuniform DIF, and crossing DIF to hold.

DDF Effects and Uniform DIF

Uniform DIF has been formally defined as a constant value of $\Lambda(\theta)$ across the θ continuum (Hanson, 1998), and this definition has been adopted for all ensuing discussions of uniform DIF. It follows that identifying the conditions required for uniform DIF to hold is tantamount to identifying the conditions required for a constant $\Lambda(\theta)$ under Equation 7. Let us assume that $a_j \neq 0$ for $j = 1, 2, \ldots, J$. Under this assumption, and the assumption that not all $\omega_j = 0$, the condition of uniform DIF will be met if, and only if, the term $\exp(-a_j\theta)$ is cancelled from Equation 7. As will be shown below, two conditions lead to the cancellation of $\exp(-a_j\theta)$ from Equation 7: (a) when the value of ω_j is constant for $j = 1, 2, \ldots, J$, and (b) when the value of a_j is constant for $j = 1, 2, \ldots, J$. Each of these two situations, in turn, is described in greater detail.

Let us first consider the situation of a constant DDF effect, such that $\omega_1 = \omega_2 = \ldots = \omega_J$. Let us denote the constant DDF effect by ω_+ (i.e., $\omega_j = \omega_+$ for $j = 1, 2, \ldots, J$). Substituting ω_+ into Equation 7 yields the following form of $\Lambda(\theta)$

$$\Lambda(\theta) = \ln\left[\frac{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_j\theta)}{\exp(-\omega_+)\sum_{j=1}^{J}\exp(-c_j)\exp(-a_j\theta)}\right]. \tag{8}$$

The summation term in the numerator and denominator of Equation 8 cancel one another, yielding

$$\Lambda(\theta) = \ln\left[\frac{1}{\exp(-\omega_+)}\right] = \omega_+. \tag{9}$$

Thus, when the DDF effects are constant across all J distractors, the DIF effect is equal to the constant DDF effect, ω_+. Because $\Lambda(\theta)$ is equal to the constant ω_+ under the condition of constant DDF effects, it follows that the DIF effect must be uniform in nature under the condition of a constant DDF effect.

The second condition that necessarily leads to uniform DIF is the presence of a constant value of the a_j parameter, such that $a_j = a_+$ for $j = 1, 2, \ldots, J$. In this situation, $\Lambda(\theta)$ can be expressed as

$$\begin{aligned}\Lambda(\theta) &= \ln\left[\frac{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_+\theta)}{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_+\theta)\exp(-\omega_j)}\right]\\ &= \ln\left[\frac{\exp(-a_+\theta)\sum_{j=1}^{J}\exp(-c_j)}{\exp(-a_+\theta)\sum_{j=1}^{J}\exp(-c_j)\exp(-\omega_j)}\right]\\ &= \ln\left[\frac{\sum_{j=1}^{J}\exp(-c_j)}{\sum_{j=1}^{J}\exp(-c_j)\exp(-\omega_j)}\right].\end{aligned} \tag{10}$$

In this case, θ falls out of the equation for $\Lambda(\theta)$ indicating that $\Lambda(\theta)$ is independent of θ. Note that this result holds even if the DDF effects are nonzero and vary in magnitude and/or sign.

From Equation 10 we see that under the condition of constant a_j, $\Lambda(\theta)$ is a weighted aggregation of $\exp(-\omega_j)$, where the weights are proportional to $\exp(-c_j)$.

The results shown in Equations 8 to 10 have particular relevance to the conceptualization of DIF under the Rasch model (Rasch, 1960). Under the Rasch model, it is easily verified (Holland & Thayer, 1988) that the value of $\Lambda(\theta)$ for a particular item is equal to the constant $\Lambda = b_F - b_R$, where b_F and b_R correspond to the Rasch model item difficulty parameters for the focal and reference groups, respectively. Because Equations 8 to 10 demonstrate that the condition of a constant $\Lambda(\theta)$ can only arise under the conditions of a constant value of ω_j or a constant value of a_j, it follows that DIF effects conceptualized under the Rasch model are limited, in theory, to a very specific set of conditions. The Rasch model's conception of DIF will hold precisely only when (a) the DDF effects are equal across all distractors or (b) all distractors are equally discriminating. All other forms of invariance, as modeled by the NRM in Equations 3 and 4, will result in a form of DIF that is not fit precisely with the Rasch model. The extent to which violations of the constant DDF effect and distractor-level discrimination properties lead to substantial violations of the constancy of $\Lambda(\theta)$ is currently unknown and requires the attention of future research.

DDF Effects and Nonuniform DIF

In the previous section, two conditions were identified that necessarily lead to uniform DIF: (a) a constant DDF effect (ω_j) across all distractors and (b) a constant a_j parameter across all distractors. All other conditions will lead to either some form of nonuniform DIF or, potentially, no DIF at all. This section presents a numeric study investigating how various forms of DDF effects lead to different forms of nonuniform DIF. Two forms of nonuniform DIF are of particular interest: (a) the general nonuniform DIF effect, concerning the overall variability of $\Lambda(\theta)$ as a function of θ, and (b) the crossing DIF effect, concerning the extent to which $\Lambda(\theta)$ varies in sign as a function of θ such that the average value of $\Lambda(\theta)$ taken across the θ continuum is near zero. As noted previously, the crossing DIF effect is of particular interest because many widely used DIF detection procedures (Dorans & Kulick, 1986; Mantel & Haenszel, 1959; Shealy & Stout, 1993) are relatively insensitive to crossing DIF effects because of the cancelling of positive and negative values of the conditional DIF effects. Because the results of Equation 7 (demonstrating the relationship between the DIF effect and the DDF effects) does not indicate a finite set of conditions leading to nonuniform or crossing DIF, an investigation of how DDF effects are related to the nonuniform and crossing DIF effects was pursued using a numeric approach. The sections that follow describe a numeric study of how the nonuniform DIF effects result from different patterns of DDF effects. The results of the numeric study are intended to shed light on the DDF patterns associated with large nonuniform and crossing DIF effects.

Method

The numeric investigation of the relationship between DDF effects and the resulting nonuniform DIF effects was pursued by examining how the conditional DIF effect, $\Lambda(\theta)$, varied as a function of the DDF effects introduced into the item. This study considered three hypothetical items that were parameterized according to the NRM as given in Equations 3 and 4. Each of the items had three distractors, and the c_j and a_j parameters of these three items were as follows: (a) for Item 1, $c_1 = 0$, $c_2 = 1.0$, $c_3 = 1.0$, $a_1 = 2.0$, $a_2 = 0.5$, and $a_3 = 0.5$; (b) for Item 2, $c_1 = -1.0$, $c_2 = 3.0$, $c_3 = 2.0$, $a_1 = 3.0$, $a_2 = 1.0$, and $a_3 = 2.0$; and (c) for Item 3, $c_1 = 0$, $c_2 = 1.0$, $c_3 = -1.0$, $a_1 = 1.5$, $a_2 = 2.5$, and $a_3 = 1.5$. The parameters of these items were selected because they were representative of those reported in previous research (Penfield, 2008; Wollack, Bolt, Cohen, & Lee, 2002). For each of these three items, the magnitude of nonuniform and crossing DIF was quantified under varying patterns of DDF, as described below.

To quantify the overall measure of nonuniform and crossing DIF for each of the three studied items, the value of $\Lambda(\theta)$ was evaluated at 0.1 intervals ranging from $\theta = -4.0$ to $\theta = 4.0$ (this was done individually for each of the three studied items described above). For each of the three studied items, this resulted in evaluating $\Lambda(\theta)$ at 81 successive points along the θ continuum (i.e., $\theta = -4.0, -3.9, -3.8, \ldots, 4.0$). Let us denote the point along the θ continuum by i, such that $i = 1, 2, \ldots, 81$, and thus $\theta_1 = -4.0$, $\theta_2 = -3.9, \ldots, \theta_{81} = 4.0$. The notation $\Lambda(\theta_i)$ will be used to explicitly represent the value of $\Lambda(\theta)$ conditional on θ_i. Using this notation, the overall degree of nonuniform DIF in a particular studied item was evaluated by the standard deviation of $\Lambda(\theta_i)$ across the 81 points, given by

$$\mathrm{NON} = \sqrt{\frac{\sum_{i=1}^{81} [\Lambda(\theta_i) - \mu]^2}{81}}. \tag{11}$$

In Equation 11, μ corresponds to the mean value of $\Lambda(\theta_i)$ across the 81 points of evaluation. The approach used in quantifying the magnitude of the nonuniform DIF effect using the standard deviation of $\Lambda(\theta_i)$ has not been presented previously in the literature, but it was deemed to be the most appropriate approach for measuring nonuniform DIF in the current context. A large value of NON indicates that the conditional DIF effect varies substantially across the θ continuum, and thus nonuniform DIF exists. Conversely, as the value of NON approaches zero, the nonuniform DIF effect diminishes to zero. Although guidelines for interpreting NON do not currently exist, the nonuniform DIF effect was considered as being large when NON > 0.3, as this represented a situation for which the typical distance between $\Lambda(\theta_i)$ and μ is 0.3. Using this standard, a "typically low" value of $\Lambda(\theta_i)$ is one for which $\Lambda(\theta_i) - \mu = -0.3$, and a "typically high" value of $\Lambda(\theta_i)$ is one for which $\Lambda(\theta_i) - \mu = 0.3$. The difference between a typically high and typically low value of $\Lambda(\theta_i)$ is equal to 0.6. According to the widely adopted ETS classification scheme, a value of the Mantel–Haenszel log-odds ratio of 0.64 is considered to be a large DIF effect (Penfield & Camilli, 2007; Zieky, 1993), which can be interpreted as meaning that a value of 0.64 is a large departure from a value of zero. It follows that a difference of 0.6 between a typically high and typically low value of $\Lambda(\theta_i)$ is consistent with the definition of a large DIF effect under the ETS classification scheme.

Next, to evaluate whether a substantial crossing DIF effect exists, the absolute value of the sum of the signed values of $\Lambda(\theta_i)$ across the 81 points divided by the sum of the unsigned values of $\Lambda(\theta_i)$ across the 81 points was considered, given by

$$\mathrm{CRS} = 1 - \frac{|\sum_{i=1}^{81} \Lambda(\theta_i)|}{|\sum_{i=1}^{81} |\Lambda(\theta_i)|}. \tag{12}$$

Using the absolute value of the summation in the numerator of Equation 12 ensured that CRS is bounded between zero and unity. The value of CRS represents the proportion of the unsigned conditional DIF effects that correspond to crossing DIF effects (i.e., DIF effects having opposite sign). If CRS is near zero, then the conditional unsigned DIF effects must be similar in value to the conditional signed DIF effects, and thus the DIF effect is largely noncrossing in nature. In contrast, if CRS is near unity, then the conditional signed DIF effects must cancel across the θ continuum, suggesting that the nonuniform effect is crossing in nature. Values of CRS exceeding 0.5 were considered a large crossing DIF effect.

Although the focus of this study was the investigation of nonuniform and crossing DIF effects, for comprehensiveness a measure of uniform DIF was computed using the mean value of $\Lambda(\theta_i)$ across the 81 points, given by

$$\text{MEAN} = \frac{\sum_{i=1}^{81} \Lambda(\theta_i)}{81}. \tag{13}$$

The value of MEAN represents the mean conditional signed log-odds ratio, being similar to the Mantel–Haenszel common log-odds ratio estimator (Mantel & Haenszel, 1959), which considers a weighted sum of the estimated conditional odds ratios. Consistent with the ETS DIF classification scheme (Penfield & Camilli, 2007; Zieky, 1993), values of MEAN exceeding 0.43 will be considered moderate and values exceeding 0.64 in magnitude will be considered large.

The values of NON, CRS, and MEAN were obtained for each of the three studied items under 12 different conditions of DDF. The 12 conditions of DDF are presented in the leftmost column of Table 1. Six of the DDF conditions contained DDF effects that were equal in sign (although differing in magnitude across the distractors) and six of the DDF conditions contained DDF effects that varied in sign and/or magnitude (divergent DDF). Items and conditions having a large value of NON and/or CRS were flagged as yielding a substantial nonuniform or crossing DIF effect, and the conditions yielding such items were noted for further review. Although the value of MEAN was obtained for comprehensiveness, the focus of the study resides in the investigation of nonuniform and crossing DIF effects, and thus the discussion of the results will focus on the values of NON and CRS. As a final note to avoid confusion, it should be stressed that the values of NON, CRS, and MEAN obtained in this study were not estimations based on actual data but rather numeric approximations to theoretical DIF effect values.

Results

The results of the study are displayed in Table 1, which presents the values of the NON, CRS, and MEAN for each of the three hypothetical items under each of the 12 DDF conditions. The results revealed three informative outcomes that shed light on (a) the relationship between DDF effects and the resulting DIF effects and (b) the potential causes of nonuniform and crossing DIF effects. First, a given DDF pattern can lead to different DIF magnitudes depending on the characteristics of the items. For example, the DDF pattern introduced in Condition 7 yields considerably different results for NON, CRS, and MEAN across Items 1, 2, and 3. For Item 1, this DDF pattern resulted in relatively large values of NON and CRS, with a negligible MEAN value. For Item 2, this DDF pattern resulted in substantial NON and MEAN values but a zero CRS value. And for Item 3, this DDF pattern resulted in near-zero values for NON, CRS, and MEAN. This shifting of the DIF effect values for a given DDF pattern across the three experimental items occurred for many of the DDF patterns, but it was particularly salient for the patterns containing divergent DDF effects (i.e., DDF Conditions 7-12). This finding indicates that the relationship between the DIF effects and DDF effects is dependent on the item properties (i.e., the specific parameters of the item under the NRM), particularly when the DDF effects vary in sign and/or magnitude across the distractors.

Second, crossing DIF only existed for the divergent DDF conditions (Conditions 7-12). In all conditions for which the DDF effects were of the same sign (Conditions 1-6), the value of CRS was zero. This finding provides empirical evidence that crossing DIF will only exist in the presence of DDF effects that vary in sign across the distractors. Consistent with the first point discussed above, the magnitude of the crossing DIF effect depended on the parameterization of the item; a given pattern of DDF effects could result in a large crossing DIF effect for one

Table 1. Results of the Investigation Relating DIF Effects to DDF Effects

DDF condition	Measure	Item 1	Item 2	Item 3
1. $\omega_1 = 0.6, \omega_2 = 0, \omega_3 = 0$	NON	0.25	0.22	0.04
	CRS	0.00	0.00	0.00
	MEAN	0.30	0.42	0.10
2. $\omega_1 = 0, \omega_2 = 0.6, \omega_3 = 0$	NON	0.10	0.18	0.15
	CRS	0.00	0.00	0.00
	MEAN	0.12	0.12	0.12
3. $\omega_1 = 0, \omega_2 = 0, \omega_3 = 0.6$	NON	0.10	0.03	0.02
	CRS	0.00	0.00	0.00
	MEAN	0.12	0.03	0.30
4. $\omega_1 = 0.6, \omega_2 = 0.6, \omega_3 = 0$	NON	0.14	0.04	0.12
	CRS	0.00	0.00	0.00
	MEAN	0.43	0.55	0.23
5. $\omega_1 = 0, \omega_2 = 0.6, \omega_3 = 0.6$	NON	0.24	0.21	0.05
	CRS	0.00	0.00	0.00
	MEAN	0.27	0.16	0.45
6. $\omega_1 = 0.6, \omega_2 = 0, \omega_3 = 0.6$	NON	0.14	0.21	0.18
	CRS	0.00	0.00	0.00
	MEAN	0.43	0.45	0.44
7. $\omega_1 = -0.6, \omega_2 = 0, \omega_3 = 0 .6$	NON	0.68	0.87	0.03
	CRS	0.38	0.00	0.00
	MEAN	–0.22	–0.42	0.09
8. $\omega_1 = 0.6, \omega_2 = 0, \omega_3 = -0.6$	NON	0.46	0.26	0.62
	CRS	0.69	0.04	0.00
	MEAN	0.12	0.35	–0.31
9. $\omega_1 = 0, \omega_2 = 0.6, \omega_3 = -0.6$	NON	0.18	0.15	0.67
	CRS	0.00	0.24	0.19
	MEAN	–0.08	0.07	–0.28
10. $\omega_1 = 0.6, \omega_2 = -0.6, \omega_3 = -0.6$	NON	0.69	0.52	0.89
	CRS	0.94	0.57	0.00
	MEAN	–0.03	0.20	–0.44
11. $\omega_1 = 0.6, \omega_2 = 0.6, \omega_3 = -0.6$	NON	0.33	0.11	0.55
	CRS	0.33	0.00	0.28
	MEAN	0.21	0.48	–0.21
12. $\omega_1 = -0.6, \omega_2 = -0.6, \omega_3 = 0.6$	NON	0.83	1.12	0.33
	CRS	0.00	0.00	0.54
	MEAN	–0.40	–0.56	–0.08

DIF = differential item functioning; DDF = differential distractor functioning; NON = nonuniform DIF; CRS = crossing DIF; MEAN = mean DIF.

item (e.g., Item 1 for DDF Condition 8) but a negligible crossing DIF effect for a different item (e.g., Item 3 for DDF Condition 8). There were no observable item properties that signaled the potential for a large crossing DIF effect (i.e., different items displayed large CRS values in different conditions). Thus, the presence of DDF effects that differ in sign does not necessarily imply the presence of crossing DIF, but crossing DIF necessarily indicates the presence of DDF effects that differ in sign. This finding indicates that crossing DIF is generated by two or more distractors that are differentially attractive for different groups (e.g., one distractor is more attractive for one group, whereas a second distractor is more attractive for a second group).

Third, although the value of CRS was only nonzero in conditions for which the DDF effects varied in sign, the value of NON displayed a similar trend in that the value of NON only reached a substantial magnitude in conditions of divergent DDF effects (Conditions 7-12). For conditions with DDF effects having the same sign, the value of NON never exceeded 0.25 and was typically less than 0.20. In contrast, for conditions with DDF effects differing in sign, the value of NON typically exceeded 0.40 and at times exceeded 0.80. This finding provides evidence that large nonuniform DIF effects are also generated by the presence of DDF effects that vary in sign. Although DDF effects that vary in magnitude alone (i.e., Conditions 1-6) can generate some nonuniform effect, large nonuniform effects appear to be generated by the presence of distractors that are differentially attractive to different groups.

Practical Implications for DIF Analyses

Several practical implications follow from the results of the preceding sections. First, uniform DIF will only exist under two specific conditions: (a) the DDF effect is identical across all distractors or (b) the discrimination associated with each distractor is identical. The first of these two conditions—DIF arising from identical DDF effects—would likely occur when the cause of the DIF affects all distractors equally. This condition most likely reflects a situation in which the cause of the DIF resides in a property of the item stem or the correct option rather than a property of one or more of the distractors. For example, there may be some content embedded in the item stem that is not familiar to the disadvantaged group, or there may be some ambiguity in the correct option that is disproportionately affecting the disadvantaged group. In this instance, there would exist a higher probability of guessing for individuals of the disadvantaged group (conditional on ability), and we can assume that the guessing is largely random across the item options, thus leading to relatively equal DDF effects across the distractors. Because the presence of constant DDF effects indicates that the causal DIF factor likely resides either in the item stem or the correct option, it follows that item content investigations occurring as a result of an item being flagged because of a substantial uniform DIF effect should target the content of the item stem and the correct option.

A second implication is that the presence of nonuniform DIF that is noncrossing in form—if substantial in size—is likely attributable to DDF effects that vary in sign across the distractors. Although it is possible for nonuniform DIF effects to arise from DDF effects that vary only in magnitude (and not in sign), the results of this study indicate that large nonuniform DIF effects are likely restricted to situations in which DDF effects differ in sign across the distractors. This indicates that content review following a substantial nonuniform DDF effect should search for evidence of multiple biasing factors across the response options that have divergent effects across the groups being compared. Identifying the specific distractors responsible for the nonuniform DIF effect requires the completion of a DDF analysis, and thus it is recommended that a traditional DIF analysis be augmented with a DDF analysis to better target the specific factors responsible for the nonuniform DIF effect. Because the findings presented in this article are based on relationships established under the NRM, the estimation of DDF effects using an approach based on the NRM (Penfield, 2008), or parametric extensions of the NRM (Thissen, Steinberg, & Wainer, 1993), may be most appropriate for relating the DDF effects to the resulting DIF effects using the relationships developed in this article.

A third implication stemming from the results of this study is that crossing DIF results from the presence of DDF effects that differ in sign across two or more distractors. This suggests that multiple factors affecting two or more options are at play, and thus content review following an item being flagged because of substantial crossing DIF should search for evidence of biasing factors that have divergent DDF effects across the groups being compared. As was described above for nonuniform DIF, identifying the distractors involved in a crossing DIF effect requires the estimation of the DDF effect for each distractor.

Developing a Theory of How DIF Occurs

The findings of this study not only are relevant to the practice of DIF analyses but also provide an initial theory of how and why different forms of DIF occur. This theory asserts that (a) uniform DIF exists only when there exists DDF effects that are constant in magnitude and sign, or all distractors are equally discriminating, and (b) substantial nonuniform and crossing DIF exists only when the DDF effects vary in sign across the distractors. This theory offers a framework for conceptualizing how different forms of DIF are generated through the differential attractiveness of distractors. When all distractors are equally differentially attractive for a particular group (i.e., ω_j is identical for all J distractors), the result is a uniform DIF effect equal to the common DDF effect across all J distractors (i.e., uniform DIF effect equals the constant ω_j). Understanding the causes of uniform DIF, then, reduces to understanding how equal DDF effects arise, with possible explanations being given in the previous section. When the differential attractiveness shifts in sign across the distractors, such that one or more distractors are relatively more attractive for one group but other distractors are relatively more attractive for the second group, the result is a large nonuniform or crossing DIF effect. In this situation, understanding the causes of nonuniform and crossing DIF reduces to uncovering how a shifting differential attractiveness might occur. The specific mechanisms underlying this situation are currently unclear, but future research may be able to shed light on this using the DDF framework proposed in this article.

Although the theory presented here of how DIF occurs is in its infancy, it provides a framework for advancing our knowledge of the causes of DIF. It is true that there are numerous caveats to be explored and problems to be solved with respect to this theory (e.g., What item properties lead to equal DDF effects? What item properties lead to DDF effects that differ in sign?), but this theory is presented here as a starting point for the development of a targeted and testable model for explaining why different forms of DIF occur.

Using DIF and DDF in Concert

The results of this article, both empirical and theoretical, provide a strong argument for supplementing an evaluation of DIF with estimates of DDF effects. Several statistical approaches have been described in the literature for estimating DDF effects (Dorans, Schmitt, & Bleistein, 1992; Penfield, 2008; Thissen et al., 1993). Of these approaches, only Penfield's odds ratio approach provides estimates of DDF effects that are consistent with the NRM. It is noted that the DDF effects estimated using extensions of the NRM (Thissen et al., 1993) may yield DDF effects that are similar to those obtained under the NRM and thus may also be a suitable method for applying the findings of this study concerning the relationship between DDF effects and DIF effects. The extent to which the findings of this article generalize to the standardization method is currently unknown. Assuming appropriate DDF effect estimates are obtained, these estimates can be used to better interpret (a) the specific form of the DIF effect and (b) the specific causes of the DIF effect. The role that the DDF effect estimates will play in strengthening the DIF analysis will depend, in part, on the method used to estimate the DIF effect. Issues concerning how DDF effects can supplement a DIF analysis are discussed below.

The implementation of DDF analyses involves the same practical issues encountered in conducting DIF analyses. The stratifying variable (proxy for ability) should contain scores that are reliable, free of construct-irrelevant variance, and have a sufficiently high number of score levels so that each score contains individuals from a narrow range of ability (Lopez Rivas, Stark, & Chernyshenko, 2009; Penfield & Camilli, 2007; Shih & Wang, 2009; Wang & Su, 2004). In the case of implementing observed score DDF analyses, such as the odds ratio approach described by Penfield (2008) and the standardization approach described by Dorans et al.

(1992), the issue of sufficiency of the stratifying variable (Holland & Thayer, 1988; Shealy & Stout, 1993; Zwick, 1990) is of equal concern as it is in implementing observed score DIF evaluation methodologies that stratify according to a summated score (e.g., standardization and Mantel–Haenszel approaches). Although the observed score is sufficient for ability under the Rasch model, the same is not true under the two- and three-parameter logistic models, and thus observed score approaches may display a bias if the assumptions of the Rasch model do not hold. Despite the potential for bias of observed score DDF effect estimates under non-Rasch conditions, research has demonstrated that with a reliable stratifying variable, the bias is expected to be small enough that it would not compromise the utility of such approaches. In particular, simulation studies have shown that the odds ratio approach for evaluating DDF effects demonstrated only a slight bias with data generated using the NRM (Penfield, 2008), and observed-score odds ratio approaches for estimating DIF effects in dichotomous and polytomous items demonstrated good statistical properties (i.e., minimal bias, Type I error rates near the nominal level) when observed scores were obtained under non-Rasch conditions (Li & Stout, 1996; Penfield, 2007; Penfield & Algina, 2003; Shealy & Stout, 1993).

Let us now turn our attention to how DDF analyses can be used in concert with DIF analyses and consider first the case in which the DIF analyst uses a nonparametric DIF evaluation method yielding only a single measure of DIF, such as the Mantel–Haenszel approaches (Mantel & Haenszel, 1959), the standardization approach (Dorans & Kulick, 1986), or the SIBTEST approach (Shealy & Stout, 1993). In this situation, the DDF effect estimates can serve the dual role of (a) shedding light on whether the DIF effect is uniform, nonuniform, or potentially crossing and (b) identifying the causes of the DIF effect. If the DDF effects are found to be relatively constant, then the DIF analyst has strong evidence that the DIF effect is uniform in nature, and thus the nonparametric approach likely serves as an appropriate approach for describing the nature of the DIF effect. In this case, the potentially biasing factor resides either in the item stem or the correct option. If, however, the DDF effects vary in sign, then the analyst has evidence that the DIF effect is nonuniform, and potentially crossing. In this case, the pattern of the DDF effects can assist the analyst in targeting the particular distractors that contain the potentially biasing factor.

Let us next consider the case in which the DIF analyst uses a DIF evaluation method that yields specific indices of uniform and nonuniform DIF (e.g., an item response theory approach, or a logistic regression approach). In this situation, the DDF effect estimates can be used to help target the particular item property responsible for the DIF effect. In the presence of only a uniform DIF effect (e.g., between-group difference in the b-parameter of an IRT analysis), a constant DDF effect across all distractors provides evidence that the potentially biasing factor resides in either the item stem or the correct option. In the presence of a nonuniform DIF effect (i.e., between-group difference in the a-parameter of an IRT analysis), the pattern of DDF effects can be used to identify which distractors are responsible for the nonuniform DIF effect.

To demonstrate how DDF effects can be used to inform the causes of the DIF effects, an example is presented here from a DIF and DDF analysis conducted on an 80-item (56 of which were multiple-choice), individually administered, preschool science test given to 507 preschool students (Greenberg, Penfield, & Greenfield, 2009). Each multiple-choice item of the test consisted of a verbal question delivered by the test giver followed by the presentation of a series of three or four pictures, of which one was the correct answer and the others were distractor options. The test taker pointed to the picture that she or he felt was the correct answer to the question. DIF and DDF analyses were conducted that compared male ($n = 237$) and female ($n = 270$) test takers. Although the sample size was relatively small, it was deemed sufficient for the purposes of the current demonstration. Nonparametric approaches were used to evaluate both the DIF and DDF effects: (a) the Mantel–Haenszel common log-odds ratio (MH-LOR) was used to evaluate the uniform DIF effect and (b) Penfield's (2008) odds ratio approach was used to evaluate the

DDF effects (the log-odds ratio for the jth distractor will be denoted by D-LOR$_j$). All DIF and DDF analyses were conducted using the DIFAS computer program (Penfield, 2005).

Item 4 of the test demonstrated a large uniform DIF effect (MH-LOR $= -0.64$), and relatively large DDF effects for both of its distractors (D-LOR$_1$ $= -0.76$ and D-LOR$_2$ $= -0.59$). The constancy of the DDF effects across both distractors leads to two conclusions: (a) the DIF effect is caused by some property of the item prompt itself or a property of the correct response, and (b) the DIF effect is restricted to a uniform effect such that the nonuniform DIF effect is negligible. Thus, even in the absence of a formal test of nonuniform DIF, the constancy in the pattern of DDF effects leads to a conclusion that the DIF effect was uniform in nature. In contrast to Item 4, Item 19 demonstrated a negligible uniform DIF effect (MH-LOR $= 0.04$) and DDF effects that varied in sign across the distractors (D-LOR$_1$ $= -0.18$, LOR$_2$ $= 0.06$, and D-LOR$_3$ $= 0.39$). The divergent DDF effects indicate that there may exist a nonuniform DIF effect, suggesting that the item be flagged for review and/or subsequent analyses for a nonuniform DIF effect. The values obtained from the DDF analysis indicate that the content review should target the content of the first and third distractors.

Discussion

The past two decades have seen the development of a large arsenal of DIF-detection methodologies for both dichotomous and polytomously scored items (Penfield & Camilli, 2007). Despite the methodological advancements for the evaluation of DIF, relatively little progress has been made concerning the actual causes of DIF effects. Because identifying the potential causes of a DIF effect plays an integral role in flagging items as being biased or contributing construct-irrelevant variance, advances in methods useful for targeting the causes of an identified DIF effect has the potential to bridge a current gap in the DIF methodology literature. This article addresses this gap by demonstrating the specific patterns of DDF effects that are necessary for uniform, nonuniform, and crossing DIF. The results of this article help shed light on the potential nature of the invariance effects underlying uniform and nonuniform DIF effects and proposes practical guidelines for targeting the cause of the DIF effect under these conditions.

The derivations provided in this article demonstrated that although the magnitude of a uniform DIF effect is completely determined by the magnitude of the DDF effects (see Equations 8 and 9), the magnitude of a nonuniform or crossing DIF effect is determined by a combination of the DDF effects and other item location and discrimination properties (i.e., c_j and a_j). The results of the numeric study found no observable relationship between the particular item properties (i.e., parameters) and the magnitude of nonuniform and crossing DIF effects, and thus it remains unclear what item properties are required in order for a large crossing or nonuniform DIF effect to result from large divergent DDF effects. Future research on this topic may shed light on the item properties required for large nonuniform or crossing DIF effects to arise.

In this study, the DIF effect was modeled as a function of the DDF effects using the NRM. It is acknowledged, however, that the NRM may not be realistic in certain circumstances because it does not have any inherent mechanism for modeling guessing in multiple-choice items. Given the impact that guessing can have on the magnitude of DIF effects (Camilli & Penfield, 1997), it would be fruitful to extend the current study to include considerations of guessing. Several polytomous response models for multiple-choice items that have the capacity to model guessing have been proposed in the literature (Samejima, 1979; Thissen & Steinberg, 1984), and future research may benefit from examining how DIF effects can be expressed as a function of DDF effects under these models.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author received no financial support for the research and/or authorship of this article.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

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

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel–Haenszel statistics. *Journal of Educational Measurement, 34*, 123-139.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Dorans, N. J., & Kulick, E. M. (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., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309-319.

Finch, W. H., & French, B. F. (2006, April). *Detection of crossing differential item functioning: A comparison of four methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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

Greenberg, A. C., Penfield, R. D., & Greenfield, D. B. (2009, April). *Fighting bias with statistics: Applying DIF and DDF to detect gender differences in responses to items on a preschool science assessment*. Poster presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*, 244-253.

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

Li, H.-H., & Stout, W. (1996). A new procedure for the detection of crossing DIF. *Psychometrika, 61*, 647-677.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251-265.

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

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

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153-171). Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389-402.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research, 49*, 231-243.
Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150-151.
Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*, 187-210.
Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement, 45*, 247-269.
Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 353-370.
Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125-167). New York, NY: Elsevier.
Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.
Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report No. 79-4). Knoxville: University of Tennessee, Department of Psychology.
Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*, 67-81.
Shealy, R. T., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as we as item bias/DIF. *Psychometrika, 58*, 197-239.
Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199.
Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.
Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice items: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.
Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.
Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel–Haenszel method. *Applied Measurement in Education, 17*, 113-144.
Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*, 339-352.
Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.
Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-198.